PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Dudek
Akademia Ekonomiczna we Wrocławiu

METODA TAKSONOMII WROCŁAWSKIEJ W KLASYFIKACJI OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

Metoda taksonomii wrocławskiej, dobrze znana w wielowymiarowej analizie statystycznej, jest obecnie trochę zapomniana, ponieważ jest podobna do metody najbliższego sąsiada klasyfikacji hierarchicznej. Okazuje się jednak, iż metoda ta może zostać z powodzeniem zastosowana w odniesieniu do danych symbolicznych (dla tych danych ze względu na specyfikę miar odległości nie pojawia się problem tożsamości algorytmu z metodą najbliższego sąsiada).

Artykuł przedstawia zastosowanie metody taksonomii wrocławskiej do klasyfikacji obiektów symbolicznych. W kolejnych częściach artykułu opisane są: pojęcie obiektu i zmiennej symbolicznej, miary podobieństwa obiektów symbolicznych, kroki metody taksonomii wrocławskiej, algorytmu klasyfikacji i mierniki jakości klasyfikacji dla danych symbolicznych, a także dwa eksperymenty pozwalające na ocenienie przydatności metody taksonomii wrocławskiej w klasyfikacji obiektów symbolicznych.

2. Obiekty i zmienne symboliczne

W przeciwieństwie do tradycyjnych danych, w których na przecięciu wiersza i kolumny znajdują się pojedyncze wartości liczbowe, poszczególne składowe obiektów symbolicznych (zmienne symboliczne) mogą być reprezentowane w postaci [Bock, Diday 2000]:

- danych numerycznych,
- łańcuchów tekstowych (danych jakościowych),

- przedziałów liczbowych,
- danych w postaci listy wartości,
- danych w postaci listy wartości wraz z wagami.

Ze względu na konstrukcję obiektów symbolicznych procedura klasyfikacyjna przebiega inaczej niż w przypadku danych liczbowych; inne są główne kroki wyboru miary odległości, wyboru algorytmu klasyfikacji i wyboru miernika jakości klasyfikacji.

3. Odległości dla obiektów symbolicznych

Struktura danych reprezentowanych w obiektach symbolicznych implikuje fakt, iż do pomiaru ich podobieństwa nie można stosować klasycznych miar takich, jak odległość miejska czy euklidesowa. Zamiast tego stosowane są inne miary. Wśród najważniejszych z nich należy wymienić miary (por. [de Carvalho, Souza 1998; Ichino, Yaguchi 1994; Malerba i in. 2001]):

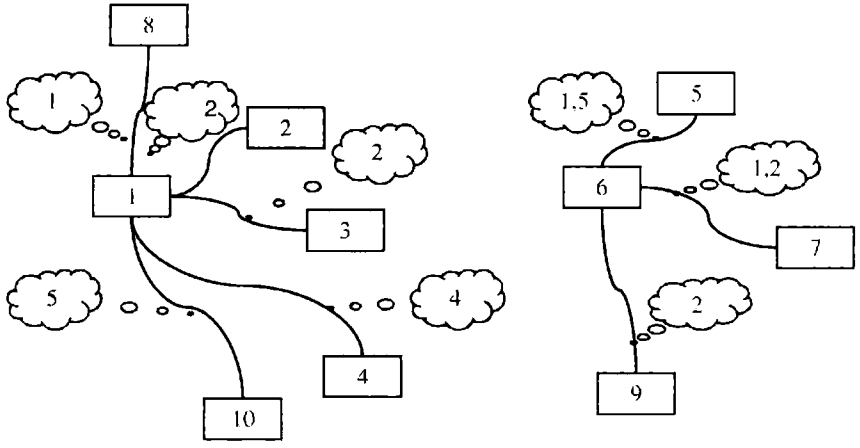
- Gowda i Krishna – miara wzajemnego sąsiedztwa (*mutual neighborhood value*),
- Hausdorffa – miara odległości między zbiorami,
- Ichino i Yaguchiego – miara oparta na pojęciach kartezjańskiego połączenia (*Cartesian join*) i kartezjańskiego przekroju (*Cartesian meet*), będących rozszerzeniami operatorów $\cup$ i $\cap$ na wszystkie typy danych reprezentowanych w obiektach symbolicznych,
- de Carvalho – rozszerzenie miar Ichino i Yaguchiego oparte na pojęciach funkcji porównującej (CF – *comparison function*) i funkcji agregującej (AF – *aggregation function*) oraz na pojęciu potencjału opisowego obiektu symbolicznego.

W odniesieniu do danych symbolicznych nie przeprowadza się zazwyczaj ich normalizacji w osobnej procedurze, natomiast odległości zarówno Ichino i Yaguchiego, jak i de Carvalho występują w wersjach połączonych z normalizacją.

4. Metoda taksonomii wrocławskiej

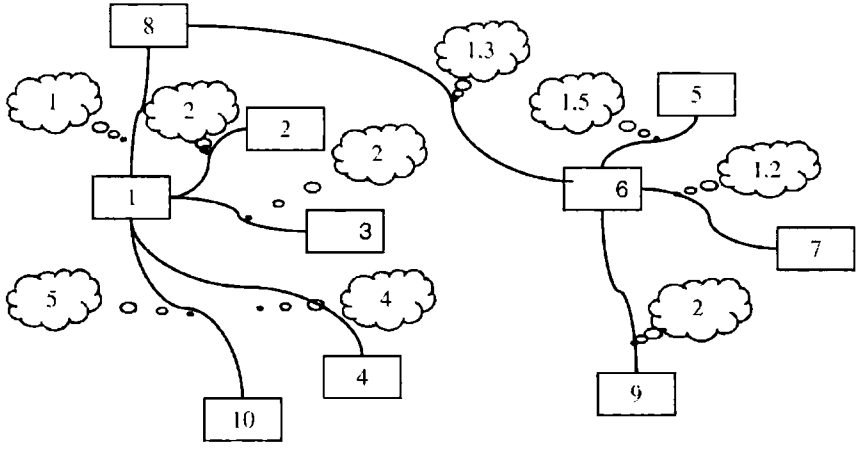
Metoda taksonomii wrocławskiej [Florek i in. 1951] jest nazywana również metodą dendrytową. Punktem wyjścia tej metody jest macierz odległości (zazwyczaj euklidesowych) pomiędzy obiektami. Klasyfikacja metodą dendrytową przebiega w trzech etapach:

Etap 1. Znalezienie dla każdego obiektu jego najbliższego sąsiada, co symbolizowane jest połączeniem w skupieniu pomiędzy tymi obiektami. W efekcie otrzymujemy graf (rys. 1), który może się składać z wielu skupień (w takim wypadku należy przejść do etapu drugiego) lub być grafem spójnym (etap drugi może zostać pominięty i należy przejść bezpośrednio do etapu trzeciego).

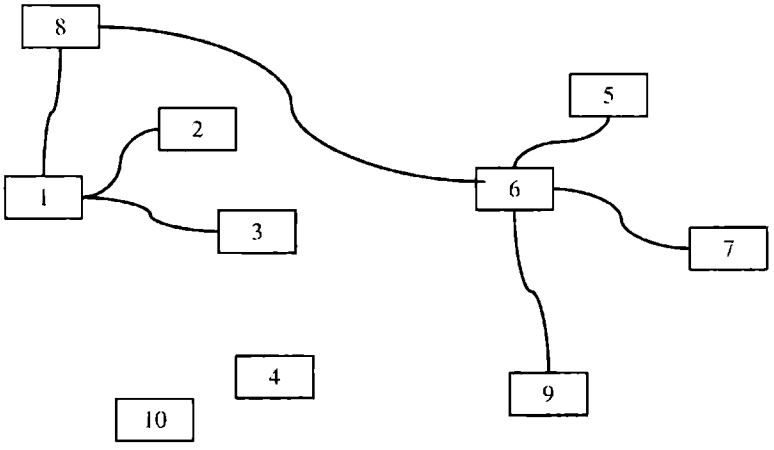


Rys. 1. Przykładowe skupienia po pierwszym etapie klasyfikacji metodą taksonomii wrocławskiej. W prostokątach znajdują się numery obiektów, a przy wiązaniach – odległości między obiektami
Źródło: opracowanie własne.

Etap 2. Łączenie skupień pierwszego rzędu w graf spójny. Dla każdego skupienia znajduje się skupienie – najbliższego sąsiada, przy czym odległość między skupieniami jest wyznaczana poprzez wyznaczenie odległości pomiędzy wszystkimi parami obiektów (O_1,O_2), z których O_1 należy do pierwszego, a O_2 do drugiego skupienia, i znalezienie najmniejszej z nich wyznaczonej przez obiekty (O_1^*,O_2^*). Obiekty O_1^* i O_2^* są następnie łączone. Etap ten jest powtarzany, aż do momentu otrzymania grafu spójnego (rys. 2).



Rys. 2. Przykładowy dendryt po drugim etapie klasyfikacji metodą taksonomii wrocławskiej. W prostokątach znajdują się numery obiektów, a przy wiązaniach – odległości między obiektami
Źródło: opracowanie własne.



Rys. 3. Przykładowe efekty klasyfikacji metodą taksonomii wrocławskiej
Źródło: opracowanie własne.

Etap 3. Graf powstały w trzecim etapie dzielimy na dendryty (odpowiadające klasom) na jeden z trzech sposobów:

a) usuwamy ustalone n-1 najdłuższych krawędzi,

b) usuwamy te krawędzie, które są dłuższe niż ustalona wartość u,

c) porządkujemy malejąco odległości w grafie i obliczamy ilorazy kolejnych odległości sąsiednich, a następnie wyznaczamy tzw. podziały naturalne.

5. Metody klasyfikacji danych symbolicznych

Podstawowym problem w stosowaniu metod klasyfikacji danych symbolicznych jest fakt, iż dla takich danych ze względu na ich charakter nie są zdefiniowane operatory dodawania i odejmowania czy pojęcie średniej. W związku z tym nie można dla nich stosować metod opartych na macierzy danych, a jedynie metody oparte na macierzy odległości.

Spośród najważniejszych takich metod wymienić należy:

a) metody hierarchiczne, aglomeracyjne (por. [Gordon 1999, s. 79]):

- metoda Warda,
- pojedynczego połączenia,
- kompletnego połączenia,
- średniej klasowej,
- ważonej średniej klasowej (Mcquitty),
- medianowa,
- środka ciężkości,

b) metody optymalizacyjne:

- metoda k-medoids [Kaufman, Rousseeuw 1990],

c) metody klasyfikacji obiektów symbolicznych (por. [Chavent i in. 2003]):

- metoda klasyfikacji obiektów symbolicznych oparta na podziałach (*divisive clustering for symbolic objects* – DIV),
- metoda klasyfikacji obiektów symbolicznych bazująca na macierzach odległości (*clustering for symbolic object based on distance tables* – DCLUST),
- metoda klasyfikacji dynamicznej obiektów symbolicznych (*dynamic clustering for symbolic objects* – SCLUST),
- metoda klasyfikacji hierarchicznej i skupiania w piramidy obiektów symbolicznych (*hierarchical & pyramidal clustering for symbolic objects* – HiPYR).

Nie można dla tego typu danych stosować popularnej metody *k*-średnich i pokrewnych takich, jak: *hard competitive learning, soft competitive learning, isodata* i innych.

Metoda taksonomii wrocławskiej może być stosowana do klasyfikacji obiektów symbolicznych, o ile do wyznaczenia macierzy odległości zostanie użyta któraś z opisanych w punkcie 3 miar odległości dla obiektów symbolicznych.

6. Ustalenie liczby klas

Podobnie jak w typowej procedurze klasyfikacyjnej, do ustalenia liczby klas dla danych symbolicznych służą mierniki jakości klasyfikacji. Z analogicznych przyczyn do przypadku wyboru metod klasyfikacji w omawianej sytuacji nie mogą być użyte wszystkie znane mierniki. Z najważniejszych indeksów, do których obliczenia używana jest macierz odległości, a nie macierz danych dla danych symbolicznych, mogą być używane trzy najlepsze mierniki globalne z eksperymentu Milligana i Cooper [1985]:

- indeks Calińskiego i Harabasza [1974],
- indeks Bakera i Huberta [Hubert 1974; Baker, Hubert 1975],
- indeks Huberta i Levine'a [1976]

oraz dwa indeksy nie uwzględnione w eksperymencie Miligana i Cooper, a dość powszechnie spotykane w literaturze:

- indeks sylwetkowy (*silhouette*) [Kaufman, Rousseeuw 1990],
- indeks Krzanowskiego i Lai (1985).

Hardy i Lellemand [2003] wśród mierników jakości klasyfikacji używanych dla danych symbolicznych wymieniają jeszcze indeks Dudy i Harta oraz indeks Beale'a.

Spośród mierników wykorzystywanych jedynie w odniesieniu do danych symbolicznych warto wymienić wskaźnik inercji symbolicznej i wskaźniki udziału (*contribution measures*) (por. [Verde, Lechevalier, Chavent 2003]).

Szerzej zagadnieniem ustalania liczby klas dla danych symbolicznych i mierników jakości klasyfikacji odnośnie do tych danych zajmuje się Dudek [2005]. Jako mierniki najbardziej odpowiednie dla danych symbolicznych wybrano mierniki

Bakera i Huberta, Huberta i Levine'a i wskaźniki udziału oparte na homogeniczności. Te indeksy wykorzystano w pierwszym eksperymencie.

7. Symulacja porównawcza

Eksperyment I (dane „sztuczne")

Wygenerowano 20 zbiorów zawierających obiekty symboliczne i dokonano klasyfikacji za pomocą dziesięciu algorytmów: metody Warda (Ward), pojedynczego połączenia (*single link*), kompletnego połączenia (*complete link*), średniej klasowej (*average*), ważonej średniej klasowej (McQuitty), medianowej (*median*), środka ciężkości (*centroid*), k-medoids (PAM), dynamicznej klasyfikacji dla obiektów symbolicznych (SCLUST) i metody dendrytowej.

Porównano mierniki jakości tych klasyfikacji. Wyniki porównania przedstawia tab. 1.

Tabela 1. Średnie mierniki jakości klasyfikacji dla wygenerowanych danych o zadanej strukturze klas

Algorytm / indeks	Baker-Hubert	Hubert-Levine	Indeks homogeniczn.
Ward	0,047	0,015	0,125
Single link	0,046	0,019	0,095
Complete link	0,046	0,021	0,063
Average	0,047	0,015	0,125
Mcquitty	0,047	0,015	0,125
Median	0,047	0,015	0,125
Centroid	0,046	0,021	0,063
PAM	0,046	0,014	0,135
SCLUST	0,036	0,009	0,072
Metoda dendrytowa	0,042	0,014	0,284

Źródło: opracowanie własne.

Można zauważyć, że dla danych sztucznie wygenerowanych metoda taksonomii wrocławskiej daje wyniki porównywalne z pozostałymi metodami, a nawet minimalnie gorsze, co może być spowodowane strukturą klas (a zwłaszcza zbliżoną liczebnością wygenerowanych klas).

Eksperyment II (dane rzeczywiste)

Dla sześciu zbiorów danych zawierających obiekty symboliczne pochodzących z repozytorium danych symbolicznych dokonano klasyfikacji za pomocą dziesięciu algorytmów (w tym metodą taksonomii wrocławskiej). Porównano zgodność wyników klasyfikacji z rzeczywistą strukturą klas za pomocą indeksu RAND i skorygowanego indeksu RAND [Hubert, Arabie 1985]. Miarę skorygowanego indeksu dla każdej procedury klasyfikacyjnej i dla każdego zbioru przedstawia tab. 2.

Tabela 2. Porównanie struktury klas za pomocą skorygowanego indeksu klas z rzeczywistą strukturą zbiorów

Algorytm / zbiór	*Car*	*Exotixicology*	*Wave*	*Web*	*Wine*
Ward	0,354	-0,140	0,297	0,011	0,452
Single link	0,354	-0,140	0,297	0,011	0,432
Complete link	0,354	-0,140	0,297	0,012	0,432
Average	0,354	-0,140	0,297	0,010	0,432
Mcquitty	0,354	-0,140	0,297	0,014	0,452
Median	0,354	-0,140	0,297	0,011	0,432
Centroid	0,354	-0,140	0,297	0,011	0,432
PAM	0,354	-0,140	0,201	0,009	0,387
SCLUST	0,113	0,160	0,389	0,015	0,476
Metoda dendrytowa	0,397	-0,001	0,318	0,021	0,483

Źródło: opracowanie własne.

Tabela 3 przedstawia średnie wartości indeksu RAND i skorygowanego indeksu RAND dla dziesięciu metod klasyfikacji.

Tabela 3. Porównanie struktury klas za pomocą skorygowanego indeksu RAND z rzeczywistą strukturą zbiorów.

Algorytm / miara	RAND	Skorygowany RAND
Ward	0,7187	0,1948
Single link	0,7124	0,1908
Complete link	0,7132	0,1910
Average	0,7122	0,1906
Mcquitty	0,7189	0,1954
Median	0,7124	0,1908
Centroid	0,7124	0,1908
PAM	0,6927	0,1622
SCLUST	0,8427	0,2306
Metoda dendrytowa	0,7357	0,2436

Źródło: opracowanie własne.

Dla rzeczywistych zbiorów danych symbolicznych (w których zazwyczaj klasy nie są równoliczne) metoda taksonomii wrocławskiej spisuje się więc dużo lepiej, dając nierzadko wyniki lepsze (bardziej zgodne z rzeczywistą strukturą) niż pozostałe metody klasyfikacji.

8. Wnioski końcowe

Metoda taksonomii wrocławskiej może być z powodzeniem stosowana dla zbiorów danych zawierających obiekty symboliczne. Metoda daje gorsze rezultaty dla zbiorów obiektów symbolicznych, w których klasy są mniej więcej równoliczne,

dla rzeczywistych zbiorów danych zawierających obiekty symboliczne może zaś dawać wyniki lepsze niż pozostałe algorytmy klasyfikacyjne.

Literatura

Baker F.B., Hubert L.J. (1975), *Measuring the Power of Hierarchical Cluster Analysis*, „Journal of the American Statistical Association" vol. 70, nr 349, s. 31-38.

Bock H.H., Diday E. (red.) (2000), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Heidelberg-Springer-Verlag.

Chavent M., De Carvalho F.A.T, Verde R., Lechevallier Y. (2003), *Trois nouvelle méthodes de classification automatique de données symboliques de type intervalle*, „Revue de Statistique Appliquée" LI 4, s. 5-29.

de Carvalho F.A.T., Souza R. (1998), *Statistical Proximity Functions of Boolean Symbolic Objects Based on Histograms*, Advances in Data Science and Classification, Heidelberg-Springer-Verlag, 391-396.

Diday E. (2002), *An Introduction to Symbolic Data Analysis and the SODAS Software*, J.S.D.A., International E-Journal.

Dudek A. (2005), *Internal Cluster Quality Indexes for Classification of Symbolic Data*, Prace Naukowe Uniwersytetu Łódzkiego, Folia Oeconomica (w druku).

Florek K., Łukasiewicz J., Perkal J., Steinhaus H., Zubrzycki S. (1951), *Taksonomia wrocławska*, „Przegląd Antropologiczny" 17, s. 193-210.

Gordon A.D. (1999), *Classification*, Chapman and Hall/CRC, London.

Hardy A., Lallemand P. (2004), *Clustering Symbolic Objects Described by Multi-Valued and Modal Variables*, [w:] D. Banks i in. (red.), *Classification, Clustering and Data Mining Applications*, Springer, Berlin, s. 325-332.

Hubert L.J. (1974), *Approximate Evaluation Technique for the Single-Link and Complete-Link Hierarchical Clustering Procedures*, „Journal of the American Statistical Association" vol. 69, nr 347, s. 698-704.

Hubert L.J., Levine J.R. (1976), *Evaluating Object Set Partitions: Free Sort Analysis and Some Generalizations*, „Journal of Verbal Learning and Verbal Behaviour" vol. 15, s. 549-570.

Ichino M., Yaguchi H. (1994), *Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis*, IEEE Transactions on Systems, Man, and Cybernetics, vol. 24, nr 4, s. 698-707.

Malerba D., Espozito F., Giovalle V., Tamma V. (2001), *Comparing Dissimilarity Measures for Symbolic Data Analysis*, materiały konferencyjne "New Techniques and Technologies for Statistcs" and "Exchange of Technology and Know-How" (ETK-NTTS'01), s. 473-481.

Milligan G.W., Cooper M.C. (1985), *An Examination of Procedures for Determining the Number of Clusters in a Data Set*, „Psychometrika" nr 2, s. 159-179.
Milligan G.W. (1996), *Clustering Validation: Results and Implications for Applied Analyses*, [w:] P. Arabie, L.J. Hubert, G. de Soete (red.), *Clustering and Classification*, World Scientific, Singapore, s. 341-375.
Verde R., Lechevalier Y., Chavent M. (2003), *Symbolic Clustering Interpretation and Visualization*, „The Electronic Journal of Symbolic Data Analysis" vol. 1, nr 1.

DENDRITE CLUSTERING IN CLASSIFICATION OF SYMBOLIC OBJECTS

Summary

Dendrite clustering is used in multivariate statistical analysis for over fifty year. Despite it has been designed for „traditional" data, after minor modifications it can be adapted to symbolic data e.g. data representing: single quantitative values, categorical values, intervals, multi-valued variables, multi-valued variables with weights.

In this paper usage of dendrite method for symbolic data is described, Nine clustering algorithms: Ward, single link, complete link, average link, Mcquitty, median and centroid hierarchical clustering methods, partitioning around medoids, dynamic clustering for symbolic objects are compared with dendrite method and situations where this method gives better results are pointed.