PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1151 2007
Ekonometria 18

Andrzej Dudek

METODY KLASYFIKACJI DLA DANYCH SYMBOLICZNYCH – SYMULACJA PORÓWNAWCZA

1. Wstęp

Jednym z najważniejszych zagadnień w procedurze klasyfikacyjnej jest wybór metody klasyfikacji. O ile zagadnienie to dla danych w postaci macierzy liczbowej jest dość dobrze opisane w literaturze, o tyle w wypadku danych symbolicznych brak jest jeszcze publikacji je poruszających. W artykule podjęta została próba porównania najważniejszych metod klasyfikacyjnych używanych dla danych symbolicznych. Przedstawiono typowe etapy procedury klasyfikacyjnej, opisano pojęcia obiektu i zmiennej symbolicznej, scharakteryzowano miary odległości dla nich używanych, przedstawiono najważniejsze metody klasyfikacji używane dla danych symbolicznych, poruszono zagadnienie ustalenia liczby klas w procedurze klasyfikacyjnej oraz zaprezentowano wyniki symulacji porównawczej dla metod klasyfikacji dla danych symbolicznych.

2. Procedura klasyfikacji – typowe etapy

W literaturze przedmiotu w typowej procedurze klasyfikacyjnej wyodrębnia się zazwyczaj osiem etapów (por. np. [17, s. 342-343]):
1) wybór obiektów do klasyfikacji,
2) wybór zmiennych charakteryzujących obiekty,
3) wybór formuły normalizacji wartości zmiennych,
4) wybór miary odległości,
5) wybór metody klasyfikacji,
6) ustalenie liczby klas,
7) walidację wyników klasyfikacji,
8) opis (interpretację) i profilowanie klas.

Uznaje się (por. [19]), że etapy 3-6 (dotyczące wyboru formuły normalizacji wartości zmiennych, miary odległości, metody klasyfikacji i ustalenia liczby klas) są newralgiczne dla procedury klasyfikacyjnej.

3. Obiekty i zmienne symboliczne

Aby opisać kolejne etapy procedury klasyfikacyjnej dla danych symbolicznych, należy wprowadzić pojęcia obiektu i zmiennej symbolicznej.

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 [1]):

- danych numerycznych,
- łańcuchów tekstowych (danych jakościowych),
- przedziałów liczbowych,
- danych w postaci listy wartości,
- danych w postaci listy wartości wraz z wagami.

Zmienne w obiektach symbolicznych mogą ponadto [6]):

- reprezentować strukturę hierarchiczną (*taxonomic variable*),
- być hierarchicznie zależne – jeśli jedna ze zmiennych jest określona tylko wtedy, gdy inna zmienna przyjmuje określoną wartość,
- być związane zależnościami logicznymi – jeśli pomiędzy wartościami zmiennych muszą zachodzić określone związki, aby dane były poprawne.

Ze względu na konstrukcję obiektów symbolicznych etapy 3-6 procedury klasyfikacyjnej różnią się dość istotnie od przypadku, gdy procedurze klasyfikacji poddawane są dane w postaci macierzy liczbowych. Te etapy, dla danych symbolicznych, zostaną omówione w dalszej części artykułu.

4. Odległości dla obiektów symbolicznych

Struktura danych reprezentowanych w obiektach symbolicznych implikuje to, 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 (por. [5; 13; 16]).

- miarę Gowdy i Krishna – miara wzajemnego sąsiedztwa (*mutual neighborhood value*),
- miarę Hausdorffa – miara odległości między zbiorami,
- miarę 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,

- miarę 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.

Dla 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. Zazwyczaj więc etapy 3 i 4 przedstawionej typowej procedury klasyfikacyjnej w przypadku danych symbolicznych łączone są w jeden etap, przy czym większość autorów zaleca, aby jako miara odległości w tych etapach wybierana była miara odległości Ichino i Yaguchiego w postaci znormalizowanej.

5. Metody klasyfikacji dla danych symbolicznych

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

Spośród najważniejszych takich metod wymienić należy:

- metody hierarchiczne, aglomeracyjne (por [9, s. 79]):
  - metodę Warda,
  - metodę pojedynczego połączenia,
  - metodę kompletnego połączenia,
  - metodę średniej klasowej,
  - metodę ważonej średniej klasowej (Mcquitty),
  - metodę medianową,
  - metodę środka ciężkości;
- metody optymalizacyjne:
  - metodę *k*-medoidów [14]);
- metody klasyfikacji obiektów symbolicznych (por. [4]):
  - metodę klasyfikacji obiektów symbolicznych opartą na podziałach (*Divisive Clustering for Symbolic Objects* – DIV),
  - metodę klasyfikacji obiektów symbolicznych bazującą na macierzach odległości (*Clustering for Symbolic Object Based on Distance Tables* – DCLUST),
  - metodę klasyfikacji dynamicznej obiektów symbolicznych (*Dynamic Clustering for Symbolic Objects* – SCLUST),
  - metodę klasyfikacji hierarchicznej i skupiania w piramidy obiektów symbolicznych (*Hierarchical & Pyramidal Clustering for Symbolic Objects* – HiPYR).

Natomiast nie można dla tego typu danych stosować popularnej metody *k*-średnich i pokrewnych, takich jak *Hard Competitive Learning*, *Soft Competitive Learning*, *Isodata* i innych.

6. Ustalenie liczby klas

Podobnie jak w typowej procedurze klasyfikacyjnej do ustalenia liczby klas dla danych symbolicznych służą mierniki jakości klasyfikacji. Z podobnych przyczyn jak przy wyborze metod klasyfikacji nie wszystkie znane mierniki mogą zostać użyte w tym wypadku. Spośród najważniejszych indeksów, do których obliczenia używana jest macierz odległości, a nie macierz danych dla danych symbolicznych, używane mogą być:

- trzy najlepsze mierniki globalne z eksperymentu Milligana i Coopera [18]:
  - indeks Calińskiego i Harabasza [3],
  - indeks Bakera i Huberta [2; 11]),
  - indeks Huberta i Levine'a [12];
- dwa indeksy nie uwzględnione w eksperymencie Milligana i Coopera, a dość powszechnie spotykane w literaturze:
  - indeks sylwetkowy (*Silhouette*) [14]),
  - indeks Krzanowskiego i Lai [15].

A. Hardy i P. Lellemand [10] wśród mierników jakości klasyfikacji używanych dla danych symbolicznych wymieniają jeszcze indeks Dudy i Harta oraz indeks Beale'a.

Spośród mierników wykorzystywanych jedynie dla danych symbolicznych warto wymienić wskaźnik inercji symbolicznej i wskaźniki udziału (*contribution measures*) (por. [18]).

Szerzej zagadnieniem ustalania liczby klas dla danych symbolicznych i mierników jakości klasyfikacji dla tych danych zajmuje się A. Dudek [8].

7. Porównanie metod klasyfikacji dla danych symbolicznych

W analizie porównano dziewięć metod klasyfikacji:

- metodę Warda (H1),
- metodę pojedynczego połączenia (H2),
- metodę kompletnego połączenia (H3),
- metodę średniej klasowej (H4),
- metodę ważonej średniej klasowej (H5),
- metodę medianową (H6),
- metodę środka ciężkości (H7),
- metodę *k*-medoidów (PAM),
- metodę klasyfikacji dynamicznej obiektów symbolicznych (SCLUST).

Z wcześniejszych badań symulacyjnych przeprowadzonych przez autora wynika, że algorytm dynamicznej klasyfikacji dla danych symbolicznych jest bardzo wrażliwy na początkowy wybór zalążków skupień, dlatego, aby zachować porównywalność w stosunku do innych metod klasyfikacji, metoda ta została w analizie uwzględniona w trzech wersjach (przy k oznaczającym liczbę klas, n – liczbę obiektów):

a) za zalążki klas uznano pierwsze k obiektów,

b) za zalążki klas uznano obiekty 1, $m+1$, $2m+1$, ..., $(k-1)m+1$ (gdzie $m=\left[\frac{n}{k}\right]$),

c) za zalążki klas posłużyły wyniki klasyfikacji otrzymane w algorytmie k-medoidów.

Analiza polegała na poddaniu klasyfikacji 105 zbiorów danych symbolicznych o znanej strukturze C_i $(i=1,2,\ldots,105)$[1] i sprawdzeniu, wyniki której z metod odzwierciedlają tę strukturę. Jeśli wyniki klasyfikacji dla określonej metody dla i-tego zbioru były całkowicie zgodne z C_i, to odpowiedni wskaźnik sukcesów dla tej metody zwiększano o 1, jeśli zaś różnica występowała tylko dla klasyfikacji jednego obiektu, wskaźnik sukcesów zwiększano o 1/2.

Jako miara odległości użyta została miara Ichino i Yaguchiego.

Obliczenia wykonane zostały w środowisku *R* z wykorzystaniem autorskiego pakietu *SymbolicDA*.

Wyniki przeprowadzonej symulacji przedstawia tab. 1. Kolumna nagłówkowa zawiera skróty nazw metod klasyfikacyjnych, wiersz nagłówkowy zaś zawiera informację o tym, ile klas powinna liczyć dokonana klasyfikacja. Dla każdej kolumny przeprowadzono piętnaście testów na różnych zbiorach i tyle wynosi maksymalna liczba sukcesów w pojedynczej komórce tabeli.

Z przeprowadzonej symulacji wynika, iż dla danych symbolicznych metodami najlepiej znajdującymi rzeczywistą strukturą klas są (w kolejności) SCLUST, metoda Warda i metoda k-medoidów.

Symulacja potwierdziła równocześnie wcześniejsze obserwacje o słabej stabilności algorytmu SCLUST względem wyboru początkowego zalążków klas, gdyż ten sam algorytm, który dawał bardzo dobre rezultaty klasyfikacji przy przyjęciu za początkowe skupienia wyników działania metody k-medoidów, spisywał się bardzo źle dla tych samych danych przy innym doborze zalążków klas.

[1] Przez zbiór o znanej strukturze rozumiany jest zbiór wygenerowany w następujący sposób: dla zmiennych liczbowych poszczególne zmienne wygenerowane są w sposób analogiczny do programu ClustGen Milligana, a dla pozostałych typów zmiennych ich wartości są tak dobrane, aby dla tej zmiennej część wspólna obiektu z jednej klasy i obiektu z innej klasy była zawsze zbiorem pustym.

Tabela 1. Wyniki symulacji porównawczej dla dziewięciu algorytmów klasyfikacji obiektów symbolicznych

Poprawne klasyfikacje/ metoda	3 klasy	4 klasy	5 klas	7 klas	9 klas	10 klas	12 klas	Razem
H1	11	12	12,5	14	11	12	12,5	85
H2	8	4,5	3	6,5	7	6,5	7	42,5
H3	7,5	10	6,5	9,5	10	8,5	8,5	60,5
H4	10	11,5	10	10	9,5	10,5	11,5	73
H5	10	11,5	10,5	11,5	10	10,5	11,5	75,5
H6	7	5,5	5	4,5	7	6	5	40
H7	5,5	7,5	4	2	5,5	6	6	36,5
PAM	12,5	12	12,5	12	11	10,5	13,5	84
SCLUST a	0	0	0	0	0	0	0	0
SCLUST b	0,5	0	0	0	0	0	0	0,5
SCLUST c	15	15	15	12	12	11	13,5	93,5

Źródło: opracowanie własne z wykorzystaniem pakietu *symbolicDA* w języku *R*.

8. Wnioski końcowe

Z przeprowadzonych badań symulacyjnych wynika, iż dla danych symbolicznych metody zaprojektowane specjalnie dla tych danych, takich jak algorytm SCLUST, dają lepsze rezultaty niż te algorytmy, które zostały stworzone dla danych „tradycyjnych", jednak dostosowano je tak, aby możliwe było ich używanie dla obiektów symbolicznych.

Algorytm klasyfikacji dynamicznej dla danych symbolicznych SCLUST daje bardzo dobre wyniki dla danych symbolicznych przy odpowiednim doborze zalążków klas. Z przeprowadzonej symulacji wynika, że najlepsze wyniki zostały osiągnięte, gdy za początkowe skupienia dla algorytmu SCLUST przyjęto klasy otrzymane w algorytmie *k*-medoidów, dlatego wydaje się, iż dla rzeczywistych problemów klasyfikacyjnych można rekomendować strategię polegającą na przeprowadzeniu dwustopniowej procedury klasyfikacyjnej łączącej metodę *k*-medoidów do utworzenia zalążków klas i metodę SCLUST do ostatecznego podziału.

Literatura

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

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

[3] Calinski R.B., Harabasz J., *A Dendrite Method for Cluster Analysis*, „Communications in Statistics" 1974, vol. 3, 1-27.

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

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

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

[7] Dudek A., *Miary podobieństwa obiektów symbolicznych, Odległość Ichino-Yaguchiego*, Ekonometria 14, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1021, 2004, 100-106.

[8] Dudek A., *Internal Cluster Quality Indexes for Clustering of Symbolic Data*, Acta Universitatis Lodziensis, Folia Oeconomica (złożone do druku).

[9] Gordon A.D., *Classification*, Chapman and Hall/CRC, London 1999.

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

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

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

[13] Ichino, M., Yaguchi H., *Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis*, „IEEE Transactions on Systems, Man, and Cybernetics" 1994, vol. 24, nr 4, 698-707.

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

[15] Krzanowski W.J., Lai Y.T. (), *A Criterion for Determining the Number of Groups in a Data Set Using Sum of Squares Clustering*, „Biometrics" 1985, 44, 23-34.

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

[17] Milligan G.W., *Clustering Validation: Results and Implications for Applied Analyses*, [w:] *Clustering and Classification*, red. P. Arabie, L.J. Hubert, G. de Soete, World Scientific, Singapore, 1996, 341-375.

[18] Milligan G.W., Cooper M.C., *An Examination of Procedures for Determining the Number of Clusters in a Data Set*, „Biometrics" 1985, 159-179.

[19] Verde R., Lechevalier Y., Chavent M., *Symbolic Clustering Interpretation and Visualization*, „The Electronic Journal of Symbolic Data Analysis" 2003, vol. 1, nr 1.

[20] Walesiak M., Dudek A., *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – oprogramowanie komputerowe i wyniki badań*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu (złożone do druku).

CLASSIFICATION METHODS FOR SYMBOLIC DATA – COMPARATIVE SIMULATION

Summary

The article presents typical cluster analysis study for symbolic data, main dissimilarity measures for symbolic data, selection of clustering methods for symbolic data and determination of the number of classes described.

Nine clustering methods: Ward, single link, complete link, average, Mcquitty, median, centroid, partitioning around medoids and dynamic clustering for symbolic objects (SCLUST) are compared on 105 sets of symbolic data and some recommendations regarding cluster analysis of symbolic data are proposed.

Dr Andrzej Dudek jest adiunktem w Katedrze Ekonometrii i Informatyki Akademii Ekonomicznej we Wrocławiu.