

Andrzej Sokołowski

Akademia Ekonomiczna w Krakowie

Krzysztof Jajuga

Akademia Ekonomiczna we Wrocławiu

IFCS-2004 CHICAGO – KILKA REFLEKSJI

1. IFCS i SKAD – kilka faktów

International Federation of Classification Societies (IFCS) to międzynarodowe towarzystwo naukowe, będące federacją narodowych, regionalnych bądź „lingwistycznych” towarzystw naukowych. Celem IFCS jest działalność naukowa, zorientowana na teorię, algorytmy i zastosowania w zakresie analizy danych i klasyfikacji. Towarzystwo powstało w 1985 roku w Cambridge (Wielka Brytania). Jego założycielami były towarzystwa naukowe obejmujące następujące obszary geograficzne: Północną Amerykę – CSNA, Wielką Brytanię – BCS, Niemcy, Austrię, Szwajcarię niemieckojęzyczną – GfKI, Japonię – JCS, Francję, Walonię, Szwajcarię francuskojęzyczną – SFC, Włochy – SIS.

Z czasem do IFCS dołączyły następujące towarzystwa naukowe: VOC z Holandii i Flandrii, SKAD z Polski, CLAD z Portugalii, KCS z Korei Południowej, IPRCS z Irlandii (również Północnej), SoCCCAD z Ameryki Środkowej i Karaibów. Przez pewien czas członkiem IFCS było również towarzystwo jugosłowiańskie (JSK), które przestało funkcjonować w połowie lat 90. Obecnie w skład IFCS wchodzi 12 towarzystw naukowych, a zainteresowanie wejściem do IFCS wyrażają towarzystwa naukowe ze Słowenii, z Grecji, Republiki Południowej Afryki oraz z Chin.

Pracami IFCS kieruje rada (IFCS Council) oraz zarząd (IFCS Executive Committee), na czele którego stoi prezydent, wybierany na dwuletnią kadencję. Godność prezydenta piastowali i nadal piastują znani profesorowie, zajmujący się analizą danych: Hans Bock (Niemcy) – 1986-1987, Robert Sokal (Stany Zjednoczone) – 1988-1989, John Gower (Wielka Brytania) – 1990-1991, William Day (Kanada) – 1992-1993, Allan Gordon (Wielka Brytania) – 1994-1995, Doug Carroll (Stany

Zjednoczone) – 1996-1997, Chikio Hayashi (Japonia) – 1998-1999, Jean Paul Ranson (Belgia) – 2000-2001, Carlo Lauro (Włochy) – 2002-2003, Henk Kiers (Holandia) – 2004-2005.

Najważniejszym wydarzeniem w działalności IFCS jest organizowana raz na dwa lata konferencja, gromadząca wszystkich czołowych przedstawicieli z dziedziny analizy danych. Do tej pory odbyło się 9 konferencji. Odbyły się one w wymienionych miastach: Akwizgran – 1987 r., Charlottesville – 1989 r., Edynburg – 1991 r., Paryż – 1993 r., Kobe – 1996 r., Rzym – 1998 r., Namur – 2000 r., Kraków – 2002 r., Chicago – 2004 r. Następna konferencja odbędzie się w 2006 r. w Lublanie.

Sekcja Klasyfikacji i Analizy Danych PTS, będąca członkiem IFCS od 1993 r., powstała w roku 1989 jako Sekcja Taksonomiczna, obecną nazwę przyjęła zaś w momencie wejścia do IFCS. Inicjatorami powstania sekcji byli profesorowie: Zdzisław Hellwig, Kazimierz Zając i Józef Pociecha. Sekcję utworzono na konferencji w Mogilanach w 1989 r. Warto dodać, że przed powstaniem sekcji ośrodek wrocławski zorganizował pięć konferencji dotyczących klasyfikacji i analizy danych. W 1979 r. zorganizowano konferencję w Szklarskiej Porębie, w 1981 r. – we Wrocławiu, w 1983 r. – w Ustroniu koło Kępna, w 1984 r. i w 1987 r. – w Zachełmiu koło Jeleniej Góry.

Konferencje regularnie są organizowane od 1993 r. Odbywały się one w miejscowościach: Karpacz, Kraków, Szklarska Poręba, Ustron, Łódź, Świeradów, Rytro, Ustron, Łądek, Międzyzdroje, Skorzęcin, Białowieża.

2. IFCS-2004 – Chicago

Dziewiąta konferencja IFCS odbyła się w Chicago w dniach 14-18 lipca 2004 r. Instytucją organizującą był Illinois Institute of Technology. Wzięło w niej udział 196 osób z 21 krajów, najwięcej uczestników przyjechało ze Stanów Zjednoczonych (79); Francji (26); Japonii (17); Niemiec (13). W konferencji wzięły udział 4 osoby z Polski i przedstawiono 3 referaty.

Ogólna liczba referatów była następująca: wygłoszono 4 referaty tzw. *keynote*, 12 referatów zaproszonych, 63 referaty prezentowane w sesjach zaproszonych oraz 91 referatów zgłoszonych przez uczestników.

Oprócz tego odbyły się 4 tzw. krótkie kursy, mające charakter tutoriali. Ciekawym wydarzeniem na zakończenie konferencji była tzw. sesja whisky, prezentowana przez Davida Wisharta, który jest ekspertem w klasyfikacji różnych gatunków szkockiej whisky.

W sumie w trakcie konferencji odbyło się 21 sesji zaproszonych. Ich tematykę podają autorzy w oryginalnym brzmieniu: *Data Analysis Issues in PLS Regression and Path Modeling, Graphics in Statistics, Clustering Methods for Multiarray Data, Latent Variables, Multiway Data Analysis, Classification and Clustering in Astronomy, Market Research, Data Mining and Web Mining Interface, Teaching and Training*

Classification and Data Analysis, Data Mining (3 sesje), Information Retrieval (2 sesje), Statistical Methods in Immunology and Medicine, Functional Magnetic Resonance Imaging, Statistical Issues in Counterterrorism, Dimension Reduction, Issues in Statistical Metrology, Data Analysis in Financial Risk Management, Memorial Session for Chikio Hayashi.

Referaty zgłoszone przez uczestników zgrupowano w 16 sesji. Były one poświęcone następującym zagadnieniom: *Approaches to Contingency Tables, Symbolic Data Analysis (2 sesje), New Approaches in Cluster Analysis, New Kinds of Problems, Graphs and Consensus, New Ideas in Classification, Text Analysis, Neural Nets and Variable Selection, Applications in Business, Issues in k-Means Clustering, Applications in Medicine, New Ideas in Cluster Analysis, New Ideas for Binary Data, Latent Variables and Related Topics, Odds and Ends.*

Analiza referatów wygłoszonych na konferencji skłania do pewnych refleksji. Jak się wydaje, referaty te odzwierciedlają pewne tendencje występujące we współczesnej teorii i praktyce analizy danych. Zdaniem autorów, tendencje te są następujące:

1. Pojawia się coraz więcej zastosowań tworzonych metod; do dziedzin, w których te zastosowania występują, zaliczyć należy przede wszystkim medycynę, marketing, finanse i analizę tekstu.

2. Przeważają dane przekrojowe, stosunkowo niewiele jest rozważań dotyczących danych w postaci szeregów czasowych.

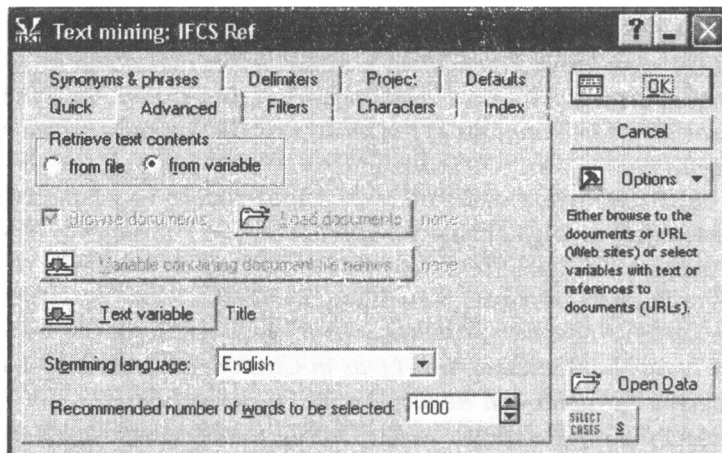
3. Występuje coraz więcej analiz nieklasycznych typów danych; do takich należy zaliczyć np. dane symboliczne, dane typu 3-way (kostka danych), dane strukturalne (*compositional*).

4. W klasyfikacji pojawia się coraz więcej metod, w których każda klasa ma u podstaw pewien model, np. rozkład wielowymiarowy.

W wielu dyskusjach prowadzonych na konferencji zwracano również uwagę na nauczanie statystycznych metod analizy danych i statystyki w ogóle. Jedną z istotnych konkluzji było to, iż nauczanie tych metod jest bardziej skuteczne, gdy powierzy się je osobie związanej z praktyką, tzn. takiej, która jednocześnie rozumie problem praktyczny i zna dostatecznie dobrze metody.

3. Analiza tytułów referatów wygłoszonych na konferencji IFCS-2004

W zbiorze streszczeń wydanych na konferencję IFCS-2004 zamieszczono krótkie opisy 167 referatów. Przygotowano plik danych zawierający tytuły tych referatów, i poddano go analizie wyszukiwania słów kluczowych za pomocą programu komputerowego *STATISTICA Text Miner*.

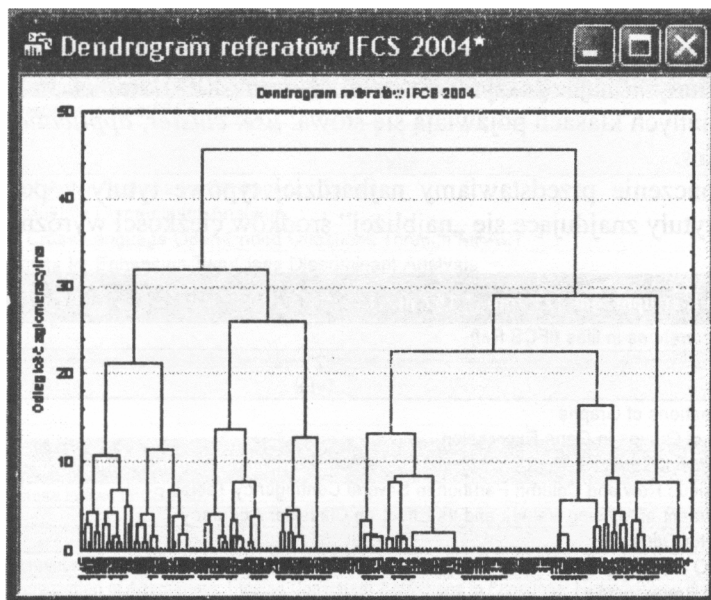


Zidentyfikowano 17 słów kluczowych. Warto zauważyć, że w procesie wyszukiwania procedura wykorzystuje rdzeń wyrazów kluczowych. Na przykład słowo kluczowe *classification* jest reprezentowane przez rdzeń *classif*. Pełną listę wyodrębnionych słów kluczowych przedstawia poniższe zestawienie.

Dane: Words summary (IFCS Ref a)

Words summary (IFCS Ref a)			
	Frequency	Number of documents	Stemmed word
algorithm	6	6	algorithm
analysis	30	29	analysi
application	8	8	applic
based	6	6	base
classification	23	22	classif
cluster	33	33	cluster
data	42	41	data
method	13	13	method
mining	7	6	mine
model	23	23	model
regression	9	8	regress
statistics	7	7	statist
symbolic	11	11	symbol
teaching	6	6	teach
trees	7	7	tree
use	13	13	use
variable	8	8	variabl

Na podstawie macierzy binarnej, kodującej występowanie słów kluczowych, przeprowadzono grupowanie tytułów za pomocą metody Warda z wykorzystaniem odległości miejskiej. Poniżej przedstawiono dendrogram.



Dendrogram ma na celu tylko ustalenie liczby klas. Zdecydowano, że zbiór referatów zostanie podzielony na cztery klasy. Ostatecznego podziału dokonano za pomocą metody *k*-średnich. Przedstawimy teraz opis klas wraz z ich proponowanymi nazwami.

Klasa 1 – „Różności”

Klasa ta obejmuje 103 referaty, co stanowi 61,7% ogółu. Najczęstsze słowa kluczowe, które tu występują to *cluster* (zawiera je 19,4% tytułów), *classification* (15,5%) oraz *model* (15,5%). W żadnym z tytułów tej klasy nie występują słowa *analysis*, *data*, *mining* oraz *use*. Częściej niż w referatach z innych klas występują tu słowa: *classification*, *method*, *algorithm*.

Klasa 2 – „Analiza danych i *data mining*”

Klasa ta liczy 37 referatów (22,2%). Najczęstsze słowa to: *data* (zawierają je wszystkie tytuły tej klasy), *analysis* (29,7%) oraz *cluster* (24,3%). Częściej niż w innych klasach występują tu słowa: *data*, *mining*, *based*, *teaching*. Tylko w tej klasie w tytułach występuje słowo *mining*.

Klasa 3 – „Metody”

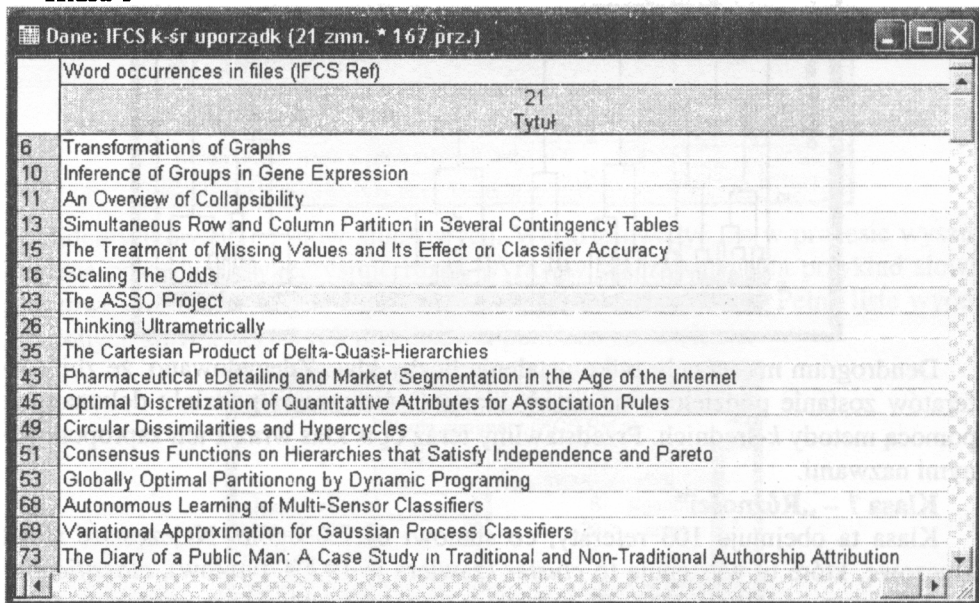
W klasie tej jest 17 referatów (10,2%). Wszystkie ich tytuły zawierają słowo *analysis*. Drugie najczęstsze słowo kluczowe to *model* (29,4%). Częściej niż w innych klasach występują tu słowa: *analysis*, *model*, *variable*, *symbolic*, *trees*. W klasie tej są referaty poświęcone różnym problemom z zakresu analizy czynnikowej, analizy dyskryminacyjnej, analizy korespondencji, drzew klasyfikacyjnych oraz metod bayesowskich.

Klasa 4 – „Zastosowania”

W klasie tej jest tylko 10 referatów (5,9%). We wszystkich tytułach występuje słowo *use*. Innymi najczęstszymi słowami są *data* (40%) oraz *cluster* (30%). Częściej niż w innych klasach pojawiają się słowa: *use*, *cluster*, *application*, *regression* oraz *statistics*.

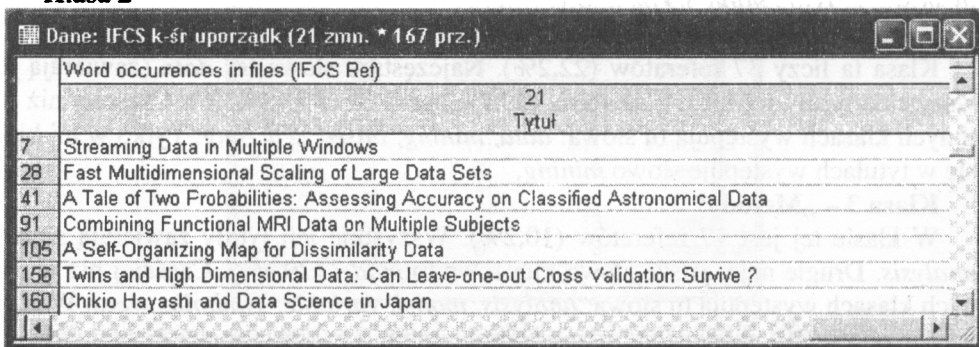
Na zakończenie przedstawiamy najbardziej typowe tytuły z poszczególnych klas. Są to tytuły znajdujące się „najbliżej” środków ciężkości wyróżnionych klas.

Klasa 1



	Word occurrences in files (IFCS Ref)	21	Tytuł
6	Transformations of Graphs		
10	Inference of Groups in Gene Expression		
11	An Overview of Collapsibility		
13	Simultaneous Row and Column Partition in Several Contingency Tables		
15	The Treatment of Missing Values and Its Effect on Classifier Accuracy		
16	Scaling The Odds		
23	The ASSO Project		
26	Thinking Ultrametrically		
35	The Cartesian Product of Delta-Quasi-Hierarchies		
43	Pharmaceutical eDetailing and Market Segmentation in the Age of the Internet		
45	Optimal Discretization of Quantitative Attributes for Association Rules		
49	Circular Dissimilarities and Hypercycles		
51	Consensus Functions on Hierarchies that Satisfy Independence and Pareto		
53	Globally Optimal Partitioning by Dynamic Programming		
68	Autonomous Learning of Multi-Sensor Classifiers		
69	Variational Approximation for Gaussian Process Classifiers		
73	The Diary of a Public Man: A Case Study in Traditional and Non-Traditional Authorship Attribution		

Klasa 2



	Word occurrences in files (IFCS Ref)	21	Tytuł
7	Streaming Data in Multiple Windows		
28	Fast Multidimensional Scaling of Large Data Sets		
41	A Tale of Two Probabilities: Assessing Accuracy on Classified Astronomical Data		
91	Combining Functional MRI Data on Multiple Subjects		
105	A Self-Organizing Map for Dissimilarity Data		
156	Twins and High Dimensional Data: Can Leave-one-out Cross Validation Survive ?		
160	Chikio Hayashi and Data Science in Japan		

Klasa 3

Dane: IFCS k-śr uporządk (21 zm., * 167 prz.)	
Word occurrences in files (IFCS Ref)	
	21 Tytuł
12	Generalized FActor Analysis for Contingency Tables
55	Sequential Multiple Discriminant Analysis
77	Analysis of Cross-Language Open-Ended Questions Through MFACT
142	Modified Biplots for Enhancing Two-Class Discriminant Analysis
158	Validation Technique in Correspondence Analysis
162	Analysis of Key Comparisons

Klasa 4

Dane: IFCS k-śr uporządk (21 zm., * 167 prz.)	
Word occurrences in files (IFCS Ref)	
	21 Tytuł
96	Customer Segmentation using gCRMi
90	Detecting a Conjunction of Alternatives: Finding Consistent Activations in Functional Neuroimaging Data using FDR
136	The Use and Ab-Use of Financial Data
154	Using Metabolomic Data to Predict Disease Status

IFCS-2004 CHICAGO – SOME IMPRESSIONS

Summary

The paper presents some facts and remarks on the world IFCS conference, which took place in Chicago. First of all, some facts on IFCS and SKAD – Polish member society are presented. Then we provide the analysis of the contents of the papers presented during conference.