Société Francophone de Classification

Bulletin numéro 22 — Décembre 2008

Présidente : C. Guinot
Vice-présidents : R. Verde et G. Venturini
Secrétaire : A. Hardy
Trésorier : J. Saracco
Éditeur : F.-X. Jollois

Société Francophone de Classification, 61 rue de Bruxelles, 5000 Namur, Belgique http://sfc.enst-bretagne.fr
Correspondance : A. Hardy, Facultés Universitaires Notre-Dame de la Paix, 8 rempart de la vierge, 5000 Namur, Belgique

Ce bulletin est le vôtre, faites-le vivre en envoyant toutes les informations que vous jugerez utiles à la communauté de la SFC, à l'éditeur en utilisant l'adresse suivante : francois-xavier.jollois@parisdescartes.fr

Sommaire

Mot de la présidente	1
Compte-rendu de l'AG	2
Parutions	4
Conférences	8
Divers	10

Mot de la présidente

L'année 2008 a été la dernière de la Présidence d'Yves Lechevallier, et cette même année Bruno Leclerc, notre Trésorier, a choisi de se retirer du Conseil d'Administration. Aussi, en tant que nouvelle présidente et au nom des représentants élus de la Société Francophone de Classification et de l'ensemble de ces membres, je tiens avant toute chose à les remercier d'avoir mis leurs compétences au service de notre association, et à les féliciter pour la qualité de leurs engagements et les résultats obtenus. Ainsi que je l'ai annoncé lors de ma candidature, mes intentions sont de mener une réflexion de fond sur les attentes et les centres d'intérêt des membres de notre association et de mettre en place les chantiers indispensables à son développement, en particulier pour faciliter l'intégration de jeunes diplômés au sein de la SFC, pour déployer une nouvelle politique de publication et de communication, et pour développer la visibilité de notre association sur le plan national et international, via notre site Internet, notre bulletin, nos publications et nos contacts.

A ma demande, le Conseil d'Administration va donc engager une évaluation complète et objective des activités menées par notre association afin de s'assurer que les mesures mises en oeuvre correspondent aux besoins concrets et aux centres d'intérêts de nos membres. Les résultats et propositions qui en découleront seront présentés en Assemblée Générale lors des prochaines Rencontres en septembre 2009 à Grenoble. Sur le plan du fonctionnement de l'association, et selon les voeux formés par notre ancien président Yves Lechevallier, le Secrétaire Général doit élaborer une modification des statuts de la SFC, afin d'élargir le nombre de représentants de notre association au sein de son Conseil d'Administration, et d'y intégrer des fonctions de Secrétaire Général-adjoint et de Trésorier-adjoint. Enfin, le Secrétaire Général sera aussi responsable de l'élaboration d'un règlement intérieur afin de préciser le fonctionnement de l'association, en particulier les domaines de responsabilité de chacun, les délégations de responsabilité, et les diverses activités de la SFC dont l'organisation des congrès.

Je souhaite à la Société Francophone de Classification de poursuivre en 2009 son expansion harmonieuse dans tous les domaines auxquels nous tenons. Le maintien de la fidélité et de l'activité de l'ensemble de nos membres, et votre soutien actif à notre association sont les garants du succès de nos activités.

Christiane Guinot
Présidente de la SFC

Paris, le 5 novembre 2008

Compte-rendu de l'Assemblée Générale de la S.F.C.

Caserta, Italie - 12 juin 2008

L'Assemblée Générale est ouverte sous la présidence de Yves Lechevallier.

1. Mot du Président de la SFC (Yves Lechevallier)

   Yves Lechevallier remercie, au nom de la SFC, toute l'équipe qui a œuvré à l'organisation et à la réussite des Rencontres conjointes SFC/CLADAG (CLassification and Data Analysis Group of the Italian Statistical Society).

2. Elections (Yves Lechevallier)

   Après dépouillement, on prend acte des résultats suivants :
   – Christiane Guinot est élue "Présidente de la SFC"
   – Jerôme Saracco est élu " Trésorier de la SFC"
   Rappelons que les mandats sont de trois ans.

   L'Assemblée remercie très chaleureusement Yves Lechevallier, président sortant, et Bruno Leclerc, trésorier, pour leur dévouement lors de leurs mandats respectifs effectués au sein de la SFC. Elle félicite Christiane Guinot et Jérôme Saracco pour leur nomination au sein du Conseil d'Administration de la SFC.

   Pour des raisons pratiques évidentes (les comptes de la SFC sont dorénavant clôturés à la fin d'une année civile), on demande à Bruno Leclerc de continuer son mandat de trésorier jusque fin 2008. Jérôme Saracco prendra donc officiellement ses fonctions de trésorier le 1 janvier 2009.

3. SFC'2008

   Les rencontres conjointes SFC/CLADAG se sont très bien déroulées. Quelques chiffres provenant des organisateurs : 138 personnes étaient présentes au congrès, dont 83 italiens, 44 francophones et 11 participants d'autres pays. Le congrès a organisé 4 sessions plénières, 4 sessions semi-plénières, 8 sessions invitées comportant 27 conférenciés invités, 16 sessions normales avec 71 papiers présentés.

   Une sélection d'articles seront publiés, d'une part en anglais dans un numéro de la série "Studies in Classification, Data Analysis, and Knowledge Organisation" et d'autre part en français dans un numéro spécial de la "Revue des Nouvelles Technologies de l'Information - C - Classification". Les personnes intéressées doivent soumettre une version longue de leur papier avant le 12 décembre 2008 à Rosanna Verde.

   Le Président remercie une fois encore très chaleureusement le comité d'organisation et le comité scientifique pour l'excellent travail fourni pour l'organisation de ces Rencontres.

4. Rapport moral (André Hardy, Secrétaire de la SFC)

   (a) Membres de la SFC

1994	1995	1996	1997	1998
63	82	63	123	116

1999	2000	2001	2002	2003
119	110	113	118	118

2004	2005	2006	2007	2008
125	100	96	94	66+...

   Tous les membres en ordre de cotisation à la SFC reçoivent les Bulletins de la SFC et les Newsletters de l'IFCS.

   (b) Bulletin de la SFC

   L'éditeur est François-Xavier Jollois. Le numéro 21 est paru en novembre 2007. Le prochain numéro paraîtra en décembre 2008.

   (c) Rencontres de la SFC

   Voici le nombre de participants aux différentes Rencontres de la SFC.

Lieu	Année	Nombres
Brest	1992	50
Tours	1994	80
Namur	1995	55
Vannes	1996	103
Lyon	1997	107
Montpellier	1998	105
Nancy	1999	82
Pointe-à-Pitre	2001	113
Toulouse	2002	120
Neuchâtel	2003	80
Bordeaux	2004	120
Montréal	2005	86
Metz	2006	70
Paris	2007	80
Caserta	2008	138 dont 44 francophones

   (d) SFC 2008

   Les 15ièmes Rencontres de la SFC se déroulent au Belvedere di S. Leucio - Caserta - Italie les 11, 12 et 13 juin 2008. Il s'agit de rencontres conjointes avec celles de la CLADAG, le groupe de Classification et d'Analyse des Données de la Société Italienne de Statistique (SIS).

   (e) SFC 2009

   Les 16èmes Rencontres de la SFC auront lieu à l'Université Joseph Fourier à Grenoble (France). Elles sont organisées par le laboratoire TIMC-IMAG. Les organisateurs locaux sont Ahlame Douzal, Gilles Bisson, Anne Guérin, Eric Gaussier, Cécile Amblard et Jérôme Gensel.

   (f) Contacts avec d'autres sociétés

   – La SFC est membre fondateur de l'IFCS ; des informations sur la SFC sont reprises dans chaque bulletin de l'IFCS ; quatre membres de la SFC participent au Conseil de l'IFCS : Bernard Fichet (élu membre additionnel), Patrice Bertrand (représentant officiel de la SFC), Jean-Paul Rasson (élu membre additionnel) et André Hardy ("Publication Officer" et membre de l'Exécutif de l'IFCS).

- La SFC a des accords privilégiés avec la Gesellschaft für Klassifikation (GfKl) et avec l'Associaçao Portuguesa de Classificaçao et Analise de Dados (CLAD).
- Une convention a été signée entre la SFC et la SFdS. Un membre de la SFC fait partie du Conseil de la SFdS (Bernard Fichet). Un membre de la SFdS fait partie du Bureau élargi de la SFC (Christian Derquenne). Une cotisation conjointe est prévue depuis 1999.
- La SFC a parrainé, pour la huitième année consécutive, les "Journées francophones d'Extraction et de Gestion des Connaissances" (Sophia-Antipolis, France, 29 janvier - 1 février 2008)
- Une collaboration scientifique a lieu entre des associations françaises et francophones (AFIA, ARIA, EGC, INFORSID, SFdS et SFC).
  - La représentante de la SFC est Pascale KUNTZ
  - Des Rencontres Inter-Associations sont régulièrement organisées.
- CSNA (Classification Society of North America) : une possibilité est donnée aux membres de la SFC de rejoindre la CSNA comme " membre affilié ", avec quelques avantages substanciels
  - Le CD "Classification Literature Automated Search Service" qui inclut un certain nombre d'ouvrages numérisés
  - L'abonnement au "Journal of Classification", et la disponibilité en ligne de la totalité des anciens numéros de ce Journal

(g) Site Web et envois électroniques

La SFC est sur Internet. Le site est pris en charge par François-Xavier Jollois. L'adresse du site : http://sfc.enst-bretagne.fr. Des liens ont été créés avec d'autres sites pouvant intéresser la SFC (IFCS, SFdS, ...). Cette année, plusieurs annonces ont été faites aux membres de la SFC par courrier électronique (congrès, annonces diverses, ...).

(h) Divers

- Habituellement, l'Assemblée Générale de la SFC se tient lors des rencontres de la SFC.
- Prochaines élections au Conseil d'Administration de la SFC : les mandats des Secrétaire et Editeur du Bulletin de la SFC arrivent à échéance. Un appel à candidature sera lancé en temps utile.

Le rapport moral est approuvé à l'unanimité.

5. Rapport financier (Bruno Leclerc)

Rapport Financier pour l'année 2007

Recettes	
Cotisations perçues durant l'année 2007	1 248,00
Régularisation de comptes avec la SFdS au titre de l'année 2005	512,00
Régularisation de comptes avec la SFdS au titre de l'année 2006	288,00
Inscriptions aux journées SFC 2007	11 240,00
TOTAL	13 288,00

Dépenses	
Prix Simon Régnier 2007	350,00
Frais bancaires (tenue de compte)	6,50
Frais pour les journées SFC 2007 (banquet, repas, cocktail)	5 881,69
TOTAL	6 238,19

Avoir de la SFC au 1er janvier 2007	10 551,54
Avoir de la SFC au 1er janvier 2008	17 601,35

Paris, le 31 août 2008,
Bruno Leclerc, trésorier de la SFC

Le rapport financier est approuvé à l'unanimité.

6. Bulletin de la SFC et site internet.

La rédaction du Bulletin de la SFC et la tenue du site web incombent à François-Xavier Jollois. François-Xavier envisage de revenir à la publication de deux numéros par an, rythme qu'il n'a pas pu assurer cette année. Il encourage les membres de la SFC à lui faire parvenir toutes les informations sur les activités pouvant intéresser les membres de la SFC : thèses, numéros spéciaux de revues, livres, workshop, congrès, ... Pour ce faire, il est possible de lui envoyer un courriel à l'adresse
francois-xavier.jollois@parisdescartes.fr

7. Cotisation SFC et cotisation conjointe SFC/SFdS

On décide de ne pas modifier les montants des cotisations à la SFC pour 2009, sauf si la SFdS modifie le montant de ses cotisations. Dans ce cas les cotisations de la SFC seraient adaptées de manière à garder le différentiel habituel de 1 euro.

8. SFC 2009

Les 16e Rencontres de la Société Francophone de Classification (SFC) se tiendront à Grenoble du 2 au 4 septembre 2009. Plusieurs équipes et laboratoires de recherche issus des universités grenobloises se mobilisent pour cette manifestation. Le programme scientifique comportera des sessions invitées et des contributions couvrant de larges thématiques connexes à la classification. Le comité de programme de cette rencontre souhaite favoriser les échanges entre chercheurs œuvrant à l'articulation de l'analyse des données et des méthodes d'apprentissage symbolique et statistique.

Les thèmes principaux des Rencontres sont

- Classification et discrimination
- Méthodes combinatoires
- Data mining
- Classifications hiérarchique et non hiérarchique
- Analyse d'image et du signal
- Approches mathématiques et statistiques
- Réseaux de neurones et algorithmes génétiques

– Représentation et visualisation
– Similarités et dissimilarités
– Analyse des données symboliques
– Arbres, graphes, treillis
– Validation
– ...

En particulier, nous souhaitons promouvoir les travaux portant sur les données suivantes
– Bio-statistique, bio-informatique
– Analyse d'information multimédia
– Analyse de données spatio-temporelles
– Apprentissage sur des données structurées (e.g. arbres, graphes)
– Analyse de données textuelles et ses applications sur le Web
– Analyse des flux de données

La première annonce de la conférence sera publiée courant décembre 2008 (http://sfc.enst-bretagne.fr/).

Le comité d'organisation.

9. SFC 2010

Jean Diatta et Henri Ralambondrainy présentent la candidature de l'Ile de la Réunion. Les Rencontres se dérouleraient sur le Campus Universitaire de Saint-Denis de la Réunion, qui est à 11 heures de vol de Paris dans l'océan indien. Le comité d'organisation provisoire comprend quatre membres. Les Rencontres pourraient avoir lieu en juin 2010. Des informations paraîtront sur le site http://tice2.univ-reunion.fr/sfc2010/.

L'assemblée approuve le choix de l'Ile de la réunion pour les Rencontres SFC'2010.

10. Accord de collaboration scientifique entre la SFC et le CLADAG

Vu le succès des Rencontres SFC/CLADAG 2008, et les liens scientifiques depuis de nombreuses années entre plusieurs membres de la Société Francophone de Classification et du Groupe Italien de Classification de la Société Italienne de Statistique, on propose de rédiger un accord de collaboration scientifique entre la SFC et le CLADAG.

Des contacts seront pris entre les présidents de la SFC et du CLADAG afin de proposer un texte et finaliser cet accord.

11. Envois des bulletins de la SFC et des Newsletters de l'IFCS

Jusqu'à présent, ces bulletins sont tout d'abord envoyés par courrier électronique. Une version papier est ensuite envoyée par courrier postal à tous les membres en ordre de cotisation.

Après divers échanges, afin d'une part de préserver la nature, et d'autre part pour éviter les coûts postaux et de reproduction, il est décidé dorénavant de ne plus envoyer ces publications que par courrier électronique. D'autre part ces publications seront toujours disponibles sur le site web de la SFC. Un courrier électronique pourra avertir les membres de la SFC de la disponibilité d'un nouveau bulletin sur le site web de la SFC

12. Prix Simon Régnier

L'Assemblée propose un montant de 450 euros pour le prix Simon Régnier en 2008.

André HARDY, Secrétaire de la SFC

Parutions

Statistical Implicative Analysis : Theory and Applications

Series : Studies in Computational Intelligence , Vol. 127
Gras, R. ; Suzuki, E. ; Guillet, F. ; Spagnolo, F. (Eds.)
2008, XVI, 514 p. 147 illus., Hardcover
ISBN : 978-3-540-78982-6

Statistical implicative analysis is a data analysis method created by Régis Gras almost thirty years ago which has a significant impact on a variety of areas ranging from pedagogical and psychological research to data mining. Statistical implicative analysis (SIA) provides a framework for evaluating the strength of implications ; such implications are formed through common knowledge acquisition techniques in any learning process, human or artificial. This new concept has developed into a unifying methodology, and has generated a powerful convergence of thought between mathematicians, statisticians, psychologists, specialists in pedagogy and last, but not least, computer scientists specialized in data mining.

This volume collects significant research contributions of several rather distinct disciplines that benefit from SIA. Contributions range from psychological and pedagogical research, bioinformatics, knowledge management, and data mining.

— — —

Algorithms for Fuzzy Clustering

Methods in c-Means Clustering with Applications
Series : Studies in Fuzziness and Soft Computing , Vol. 229
Miyamoto, Sadaaki, Ichihashi, Hidetomo, Honda, Katsuhiro
2008, XII, 248 p. 62 illus. With online files/update., Hardcover
ISBN : 978-3-540-78736-5

The main subject of this book is the fuzzy c-means proposed by Dunn and Bezdek and their variations including recent studies. A main reason why we concentrate on fuzzy c-means is that most methodology and application studies in fuzzy clustering use fuzzy c-means, and hence fuzzy c-means should be considered to be a major technique of clustering in general, regardless whether one is interested in fuzzy methods or not. Unlike most studies in fuzzy c-means, what we emphasize in this book is a family of algorithms using entropy or entropy-regularized methods which are less known, but we consider the entropy-based method to be another useful method of fuzzy c-means. Throughout this book one of our intentions is to uncover theoretical and methodological differences between the

Dunn and Bezdek traditional method and the entropy-based method. We do note claim that the entropy-based method is better than the traditional method, but we believe that the methods of fuzzy c-means become complete by adding the entropy-based method to the method by Dunn and Bezdek, since we can observe natures of the both methods more deeply by contrasting these two.

— — —

Spectral Clustering, Ordering and Ranking

Statistical Learning with Matrix Factorizations
Ding, Chris, Zha, Hongyuan
2008, Approx. 400 p. 60 illus., 20 in color., Hardcover
ISBN : 978-0-387-30448-9
Due : June 2009

Data mining methods are essential for analyzing the ever-growing massive quantities of data. Data clustering is one of the key data mining techniques. In recent years, spectral clustering has been developed as an effective approach to data clustering. It starts with well-motivated objective functions ; optimization eventually leads to eigenvectors as the solutions, with many clear and interesting algebraic properties.

Spectral clustering, ordering and ranking extensively uses matrix-based methods and algorithms. This approach is amenable to vigorous analysis and is benefiting from a treasury of knowledge of linear algebra and applied mathematics accumulated through the centuries. This exposition presents recent advances in this new subfield. New concepts are carefully developed and practical examples are extensively utilized to illustrate the ideas. A key feature are the mathematical proofs outlined throughout the text in reasonable detail which highlight the rigorous and principled quality of spectral clustering. A concise introduction to data clustering methods is followed by advanced spectral clustering, ordering and ranking topics which then lead to applications in web and text mining and genomics. An Appendix covering the preliminaries makes this text self-contained.

This book is aimed at senior undergraduate and graduate students in computer science, applied mathematics and statistics and researchers and practitioners in machine learning, data mining, multivariate statistics, matrix computation, web analysis, text mining, bioinformatics.

— — —

Modern Multivariate Statistical Techniques

Regression, Classification, and Manifold Learning
Series : Springer Texts in Statistics
Izenman, Alan Julian
2008, XXVI, 734 p., Hardcover
ISBN : 978-0-387-78188-4

Remarkable advances in computation and data storage and the ready availability of huge data sets have been the keys to the growth of the new disciplines of data mining and machine learning, while the enormous success of the Human Genome Project has opened up the field of bioinformatics.

These exciting developments, which led to the introduction of many innovative statistical tools for high-dimensional data analysis, are described here in detail. The author takes a broad perspective ; for the first time in a book on multivariate analysis, nonlinear methods are discussed in detail as well as linear methods. Techniques covered range from traditional multivariate methods, such as multiple regression, principal components, canonical variates, linear discriminant analysis, factor analysis, clustering, multidimensional scaling, and correspondence analysis, to the newer methods of density estimation, projection pursuit, neural networks, multivariate reduced-rank regression, nonlinear manifold learning, bagging, boosting, random forests, independent component analysis, support vector machines, and classification and regression trees. Another unique feature of this book is the discussion of database management systems.

This book is appropriate for advanced undergraduate students, graduate students, and researchers in statistics, computer science, artificial intelligence, psychology, cognitive sciences, business, medicine, bioinformatics, and engineering. Familiarity with multivariable calculus, linear algebra, and probability and statistics is required. The book presents a carefully-integrated mixture of theory and applications, and of classical and modern multivariate statistical techniques, including Bayesian methods. There are over 60 interesting data sets used as examples in the book, over 200 exercises, and many color illustrations and photographs.

Alan J. Izenman is Professor of Statistics and Director of the Center for Statistical and Information Science at Temple University. He has also been on the faculties of Tel-Aviv University and Colorado State University, and has held visiting appointments at the University of Chicago, the University of Minnesota, Stanford University, and the University of Edinburgh. He served as Program Director of Statistics and Probability at the National Science Foundation and was Program Chair of the 2007 Interface Symposium on Computer Science and Statistics with conference theme of Systems Biology. He is a Fellow of the American Statistical Association.

— — —

Data Mining : Foundations and Practice

Series : Studies in Computational Intelligence , Vol. 118
Lin, T.Y. ; Xie, Y. ; Wasilewska, A. ; Liau, C.-J. (Eds.)
2008, XVI, 562 p. 129 illus., 25 in color., Hardcover
ISBN : 978-3-540-78487-6

This book contains valuable studies in data mining from both foundational and practical perspectives. The foundational studies of data mining may help to lay a solid foundation for data mining as a scientific discipline, while the practical studies of data mining may lead to new data mining paradigms and algorithms.

The foundational studies contained in this book focus on a broad range of subjects, including conceptual framework of data mining, data preprocessing and data mining as generalization, probability theory perspective on fuzzy systems, rough set methodology on missing values, inexact

multiple-grained causal complexes, complexity of the privacy problem, logical framework for template creation and information extraction, classes of association rules, pseudo statistical independence in a contingency table, and role of sample size and determinants in granularity of contingency matrix.

The practical studies contained in this book cover different fields of data mining, including rule mining, classification, clustering, text mining, Web mining, data stream mining, time series analysis, privacy preservation mining, fuzzy data mining, ensemble approaches, and kernel based approaches.

We believe that the works presented in this book will encourage the study of data mining as a scientific field and spark collaboration among researchers and practitioners.

— — —

Learning Classifier Systems in Data Mining

Series : Studies in Computational Intelligence , Vol. 125
Bull, Larry; Ester, Bernadó-Mansilla; Holmes, John (Eds.)
2008, X, 230 p. 65 illus., Hardcover
ISBN : 978-3-540-78978-9

Just over thirty years after Holland first presented the outline for Learning Classifier System paradigm, the ability of LCS to solve complex real-world problems is becoming clear. In particular, their capability for rule induction in data mining has sparked renewed interest in LCS. This book brings together work by a number of individuals who are demonstrating their good performance in a variety of domains.

The first contribution is arranged as follows : Firstly, the main forms of LCS are described in some detail. A number of historical uses of LCS in data mining are then reviewed before an overview of the rest of the volume is presented. The rest of this book describes recent research on the use of LCS in the main areas of machine learning data mining : classification, clustering, time-series and numerical prediction, feature selection, ensembles, and knowledge discovery.

— — —

Computational Intelligence

Methods and Techniques
Rutkowski, Leszek
2008, XIV, 514 p. 242 illus., Hardcover
ISBN : 978-3-540-76287-4

This book focuses on various techniques of computational intelligence, both single ones and those which form hybrid methods. Those techniques are today commonly applied issues of artificial intelligence, e.g. to process speech and natural language, build expert systems and robots. The first part of the book presents methods of knowledge representation using different techniques, namely the rough sets, type-1 fuzzy sets and type-2 fuzzy sets. Next various neural network architectures are presented and their learning algorithms are derived. Moreover, the family of evolutionary algorithms is discussed, in particular the classical genetic algorithm, evolutionary strategies and genetic programming, including connections between these techniques and neural networks and fuzzy systems. In the last part of the book, various methods of data partitioning and algorithms of automatic data clustering are given and new neuro-fuzzy architectures are studied and compared. This well-organized modern approach to methods and techniques of intelligent calculations includes examples and exercises in each chapter and a preface by Jacek Zurada, president of IEEE Computational Intelligence Society (2004-05).

— — —

Applied Pattern Recognition

Series : Studies in Computational Intelligence , Vol. 91
Bunke, Horst; Kandel, Abraham; Last, Mark (Eds.)
2008, XII, 246 p. 110 illus., 51 in color. With online files/update., Hardcover
ISBN : 978-3-540-76830-2

A sharp increase in the computing power of modern computers, accompanied by a decrease in the data storage costs, has triggered the development of extremely powerful algorithms that can analyze complex patterns in large amounts of data within a very short period of time. Consequently, it has become possible to apply pattern recognition techniques to new tasks characterized by tight real-time requirements (e.g., person identification) and/or high complexity of raw data (e.g., clustering trajectories of mobile objects). The main goal of this book is to cover some of the latest application domains of pattern recognition while presenting novel techniques that have been developed or customized in those domains.

— — —

Survey of Text Mining II

Clustering, Classification, and Retrieval
Berry, Michael W.; Castellanos, Malu (Eds.)
2008, XVI, 240 p. 55 illus., Hardcover
ISBN : 978-1-84800-045-2

The proliferation of digital computing devices and their use in communication has resulted in an increased demand for systems and algorithms capable of mining textual data. Thus, the development of techniques for mining unstructured, semi-structured, and fully-structured textual data has become increasingly important in both academia and industry.

This second volume continues to survey the evolving field of text mining - the application of techniques of machine learning, in conjunction with natural language processing, information extraction and algebraic/mathematical approaches, to computational information retrieval. Numerous diverse issues are addressed, ranging from the development of new learning approaches to novel document clustering algorithms, collectively spanning several major topic areas in text mining.

Features :

– Acts as an important benchmark in the development of current and future approaches to mining textual information

- Serves as an excellent companion text for courses in text and data mining, information retrieval and computational statistics
- Experts from academia and industry share their experiences in solving large-scale retrieval and classification problems
- Presents an overview of current methods and software for text mining
- Highlights open research questions in document categorization and clustering, and trend detection
- Describes new application problems in areas such as email surveillance and anomaly detection

Survey of Text Mining II offers a broad selection in state-of-the art algorithms and software for text mining from both academic and industrial perspectives, to generate interest and insight into the state of the field. This book will be an indispensable resource for researchers, practitioners, and professionals involved in information retrieval, computational statistics, and data mining.

— — —

Machine Learning for Audio, Image and Video Analysis

Theory and Applications
Series : Advanced Information and Knowledge Processing
Camastra, Francesco, Vinciarelli, Alessandro
2008, XVI, 496 p. 99 illus. With online files/update., Hardcover
ISBN : 978-1-84800-006-3

Machine Learning involves several scientific domains including mathematics, computer science, statistics and biology, and is an approach that enables computers to automatically learn from data. Focusing on complex media and how to convert raw data into useful information, this book offers both introductory and advanced material in the combined fields of machine learning and image/video processing.

The machine learning techniques presented enable readers to address many real world problems involving complex data. Examples covering areas such as automatic speech and handwriting transcription, automatic face recognition, and semantic video segmentation are included, along with detailed introductions to algorithms and examples of their applications.

The book is organized in four parts : The first focuses on technical aspects, basic mathematical notions and elementary machine learning techniques. The second provides an extensive survey of most relevant machine learning techniques for media processing, while the third part focuses on applications and shows how techniques are applied in actual problems. The fourth part contains detailed appendices that provide notions about the main mathematical instruments used throughout the text.

Students and researchers needing a solid foundation or reference, and practitioners interested in discovering more about the state-of-the-art will find this book invaluable. Examples and problems are based on data and software packages publicly available on the web.

— — —

Support Vector Machines

Series : Information Science and Statistics
Steinwart, Ingo, Christmann, Andreas
2008, XVI, 602 p., 25 illus., 2 in color., Hardcover
ISBN : 978-0-387-77241-7

This book explains the principles that make support vector machines (SVMs) a successful modelling and prediction tool for a variety of applications. The authors present the basic ideas of SVMs together with the latest developments and current research questions in a unified style. They identify three reasons for the success of SVMs : their ability to learn well with only a very small number of free parameters, their robustness against several types of model violations and outliers, and their computational efficiency compared to several other methods.

Since their appearance in the early nineties, support vector machines and related kernel-based methods have been successfully applied in diverse fields of application such as bioinformatics, fraud detection, construction of insurance tariffs, direct marketing, and data and text mining. As a consequence, SVMs now play an important role in statistical machine learning and are used not only by statisticians, mathematicians, and computer scientists, but also by engineers and data analysts.

The book provides a unique in-depth treatment of both fundamental and recent material on SVMs that so far has been scattered in the literature. The book can thus serve as both a basis for graduate courses and an introduction for statisticians, mathematicians, and computer scientists. It further provides a valuable reference for researchers working in the field.

The book covers all important topics concerning support vector machines such as : loss functions and their role in the learning process ; reproducing kernel Hilbert spaces and their properties ; a thorough statistical analysis that uses both traditional uniform bounds and more advanced localized techniques based on Rademacher averages and Talagrand's inequality ; a detailed treatment of classification and regression ; a detailed robustness analysis ; and a description of some of the most recent implementation techniques. To make the book self-contained, an extensive appendix is added which provides the reader with the necessary background from statistics, probability theory, functional analysis, convex analysis, and topology.

— — —

Revue ADAC

Advances in Data Analysis and Classification

Volume 2, Number 1 / avril 2008
Éditeur Springer Berlin / Heidelberg
ISSN 1862-5347 (Print) 1862-5355 (Online)

- Editorial, Hans-Hermann Bock
- Two local dissimilarity measures for weighted graphs with application to protein interaction networks, Jean-Baptiste Angelelli, Anaïs Baudot, Christine Brun and Alain Guénoche

- SVM-Maj : a majorization approach to linear support vector machines with different hinge errors, P. J. F. Groenen, G. Nalbantov and J. C. Bioch
- Plastic card fraud detection using peer group analysis, David J. Weston, David J. Hand, Niall M. Adams, Christopher Whitrow and Piotr Juszczak
- Interpolation of spatial and spatio-temporal Gaussian fields using Gaussian Markov random fields, L. Fontanella, L. Ippoliti, R. J. Martin and S. Trivisonno
- Generalized constrained co-inertia analysis, Pietro Amenta

Volume 2, Number 2 / octobre 2008
Éditeur Springer Berlin / Heidelberg
ISSN 1862-5347 (Print) 1862-5355 (Online)

- Editorial, Hans-Hermann Bock, Wolfgang Gaul, Akinori Okada and Maurizio Vichi
- On multi-way metricity, minimality and diagonal planes, Matthijs J. Warrens
- Fitting semiparametric clustering models to dissimilarity data, Maurizio Vichi
- Cluster analysis of census data using the symbolic data approach, Antonio Giusti and Laura Grassini
- Multiple taxicab correspondence analysis, V. Choulakian

Conférences

EGC 2009

Extraction et Gestion de Connaissances
27–30 Janvier 2009
Strasbourg, France

Dans le prolongement des huit éditions précédentes, EGC 2009 ambitionne de regrouper chercheurs, industriels et utilisateurs francophones issus des communautés Bases de Données, Apprentissage, Représentation des Connaissances, Gestion de Connaissances, Statistique et Fouille de données. Aujourd'hui, de grandes masses de données structurées ou semi-structurées sont accessibles dans les bases de données d'entreprises ainsi que sur la toile. Aussi les entreprises ont-elles besoin de méthodes et d'outils capables de les acquérir, de les stocker, de les représenter, de les indexer, de les intégrer, de les classifier, d'extraire les connaissances pertinentes pour les décideurs et de les visualiser. Pour répondre à cette attente, de nombreux projets de recherche se développent autour de l'extraction de connaissances à partir de données (Knowledge Discovery in Data), ainsi que sur la gestion de connaissances (Knowledge Management).

L'objectif de ces journées est de rassembler, d'une part les chercheurs des disciplines connexes (apprentissage, statistique et analyse de données, systèmes d'information et bases de données, ingénierie des connaissances, etc.), et d'autre part les spécialistes d'entreprises qui déploient des méthodes d'extraction et de gestion des connaissances, afin de contribuer à la formation d'une communauté scientifique dans le monde francophone autour de cette double thématique de l'extraction et de la gestion de connaissances.

Les thématiques du colloque se regroupent en cinq catégories :

- Acquisition, recueil, pré-traitement, filtrage et fusion de données et/ou de connaissances
- Différents types de données (données spatiales, temporelles, complexes...)
- Théories, méthodes et algorithmes
- Post-traitement et modélisation de la fouille et des connaissances
- Applications innovantes

Par ailleurs, cette année, un accent tout particulier sera mis sur quatre aspects particuliers de l'extraction et de la gestion des connaissances :

- Données spatiales et temporelles
- Etude diachronique des connaissances à partir de données évolutives, avec, en particulier, les applications à la fouille d'opinions
- Profilage des utilisateurs et la fouille de données respectant la propriété intellectuelle et la vie privée
- Fouille de données web et les applications au web sémantique
- Fusion de données

http://lsiit.u-strasbg.fr/egc09/

— — —

SLDS 2009

Symposium on Learning and Data Science
1–3 avril 2009
Paris Dauphine, France

Great progress has been made in the past 20 years in Machine Learning and Statistical Learning, Data Analysis and Data Mining. From the statistical analysis of data to data mining, from machine learning to knowledge discovery, the development of data exploration and modeling has overcome numerous challenges and has benefited greatly from varied, often overlapping, paradigms.

By uniting specialists with different expertise and from different disciplines, the objectives are to compare approaches to data, to deepen understanding of different methodologies, and to focus on the Grand Challenges that must be addressed in the coming years.

The objectives of this Symposium will be achieved through presentations on core and also innovative themes. The precise topics will be chosen by the invited speakers, who have been the founders or the developers of their selected theme areas in recent years. The presentation and discussant format chosen will facilitate discussion with all of the participants.

This Symposium is important for researchers and all who want to keep abreast of the latest developments in data handling. In addition there is a competition specifically addressed to young researchers.

Award submission deadline : January 30, 2009
Post-symposium proceedings : May 31, 2009
http://ceremade.communication-pro.fr/

— — —

IFCS 2009

International Federation of Classification Societies 2009 Conference
13–18 mars 2009
Dresden University of Technology, Allemagne

The 11th Biennial Conference of the International Federation of Classification Societies (IFCS) will take place at the University of Technology of Dresden, Germany, March 13–18, 2009, in combination with the 33rd annual conference of the German Classification Society – Gesellschaft für Klassifikation (GfKl). The conference will be hosted by the Faculty of Business and Management of Dresden University of Technology in Dresden (Auditorium Centre).

The IFCS is a non-profit and non-political scientific organization which promotes the dissemination of technical and scientific information concerning data analysis, classification, related methods, and their applications. The IFCS encourages young researchers to attend its conferences, and special arrangements will be offered to them for the IFCS-2009 in Dresden.

Post-conference special issues deadline : April 30th, 2009

http://www.ifcs2009.de/

— — —

ECDM 2009

European Conference on Data Mining
18–20 Juin 2009
Algarve, Portugal

The European Conference on Data Mining (ECDM'09) is aimed to gather researchers and application developers from a wide range of data mining related areas such as statistics, computational intelligence, pattern recognition, databases and visualization. ECDM'09 is aimed to advance the state of the art in data mining field and its various real world applications. ECDM'09 will provide opportunities for technical collaboration among data mining and machine learning researchers around the globe.

The best paper authors will be invited to publish extended versions of their papers in the IADIS Journal on Computer Science and Information Systems (ISSN : 1646-3692)

Submission deadline : 30 January 2009

http://www.datamining-conf.org/

— — —

KDD 2009

Knowledge Discovery in Databases
28 Juin–1er Juillet 2009
Paris, France

The annual ACM SIGKDD conference is the premier international forum for data mining researchers and practitioners from academia, industry, and government to share their ideas, research results and experiences. KDD-09 will feature keynote presentations, oral paper presentations, poster sessions, workshops, tutorials, panels, exhibits, demonstrations, and the KDD Cup competition.

Electronic Paper Submission : February 6, 2009

http://www.sigkdd.org/kdd2009/index.html

— — —

SFC 2009

XVIème Rencontres de la Société Francophone de Classification
2–4 septembre 2009
Grenoble, France

site en préparation

— — —

ECML/PKDD 2009

European Conference on Machine Learning
Principles and Practice of Knowledge Discovery in Databases
7–11 Septembre 2009
Bled, Slovenia

The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) will take place in Bled, Slovenia, from September 7th to 11th, 2009. This event builds upon a very successful series of 19 ECML and 12 PKDD conferences, which have been jointly organized for the past eight years. It has become the major European scientific event in these fields and in 2009 it will comprise presentations of contributed papers and invited speakers, a wide program of workshops and tutorials, a discovery challenge, a demo track and an industrial track.

Paper Submission due April 20th

http://www.ecmlpkdd2009.net/

— — —

CLADAG 2009

VII riunione scientifica del CLADAG
9–11 Septembre 2009

The seventh biennial conference of the CLAssification and DAta Analysis Group (CLADAG) of the Italian Statistical Society will be held at the University of Catania, September 2009. Researchers interested in classification, cluster analysis, data analysis, multivariate analysis, computational statistics and their applications are invited to participate.

http://dssm.unipa.it/cladag/

— — —

Discovery Science 2009

3–5 Octobre 2009
Porto, Portugal

The Twelfth International Conference on Discovery Science (DS09) will be held in Porto, Portugal, on 3-5 October 2009. DS09 will be collocated with ALT09, the Twentieth International Conference on Algorithmic Learning Theory. The two conferences will be held in parallel, and will share their invited talks. The proceedings of DS09 will appear in the Lecture Notes in Artificial Intelligence Series by Springer-Verlag.

DS09 provides an open forum for intensive discussions and exchange of new ideas among researchers working in the area of Discovery Science. The scope of the conference includes the development and analysis of methods for automatic scientific knowledge discovery, machine learning, intelligent data analysis, theory of learning, as well as their application to knowledge discovery. Very welcome are papers that focus on dynamic and evolving data, models and structures.

Submission deadline : 10 May 2009

http://www.liaad.up.pt/~ds09/

Divers

Packages R

Voici une liste de packages orientés classification disponibles dans le logiciel R, et dont une nouvelle version vient de sortir.

flexmix	Flexible Mixture Modeling
mclust	Model-Based Clustering Normal Mixture Modeling
hybridHclust	Hybrid hierarchical clustering
flexclust	Flexible Cluster Algorithms
bayesm	Bayesian Inference for Marketing/Micro-econometrics
mixtools	Tools for analyzing mixture models

www.R-project.org

— — —

MIXMOD 2.1

Une nouvelle version de MIXMOD vient de sortir. Elle inclue en plus les fonctionnalités suivantes

- Nouveaux modèles pour les données de grande dimension
- Boîtes à outils Matlab et Scilab améliorées
- Méthodes d'initilisation améliorées (SMALL_EM, CEM_INIT, SEM_MAX et USER_PARTITION)
- Procédure d'installation simplifiée

Pour rappel, MIXMOD (MIXture MODelling) est un logiciel permettant de résoudre des problèmes d'estimation de densités, de classification et d'analyse discriminante. Une large variété d'algorithmes est proposé (EM, CEM, SEM) pour maximiser la vraisemblance ou la vraisemblance complétée. Le logiciel est utilisable sur des données quantitatives (modèles de mélange Gaussien) ou qualitatives (modèles de mélange Multinomial). De plus, des critères (BIC, ICL, NEC, CV) vous permettent de choisir le meilleur modèle. MIXMOD est interfacé avec Scilab et Matlab, et est distribué avec la licence GNU General Public License (GPL).

http://www-math.univ-fcomte.fr/mixmod/

— — —

TANAGRA

Nouveautés TANAGRA – 26 octobre 2008 – version 1.4.28

Dans le cadre d'un didactciel consacré à la comparaison de plusieurs logiciels libres lors de la mise en oeuvre de la méthode des centres mobiles (K-Means), les sorties du composant K-Means (et par extension, les composants dédiés à la classification automatique sur variables continues) ont été améliorées.

Pour rappel, TANAGRA est un logiciel gratuit de DATA MINING destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

TANAGRA est un projet ouvert au sens qu'il est possible à tout chercheur d'accéder au code et d'ajouter ses propres algorithmes pour peu qu'il respecte la licence de distribution du logiciel.

L'objectif principal du projet TANAGRA est d ?offrir aux chercheurs et aux étudiants une plate-forme de Data Mining facile d'accès, respectant les standards des logiciels du domaine, notamment en matière d'interface et de mode de fonctionnement, et permettant de mener des études sur des données réelles et/ou synthétiques.

Le second objectif de TANAGRA est de proposer aux chercheurs une architecture leur permettant d'implémenter aisément les techniques qu'ils veulent étudier, de comparer les performances des algorithmes. TANAGRA se comporte plus comme une plate-forme d'expérimentation qui leur permettrait d'aller à l'essentiel en leur épargnant toute la partie ingrate de la programmation de ce type d'outil : la gestion des données.

Le troisième et dernier objectif, en destination des apprentis programmeurs, vise à diffuser une méthodologie possible d'élaboration de ce type de logiciel. L'accès au code leur permettra de voir comment se construit ce type de logiciel, quels sont les écueils à éviter, quelles sont les principales étapes d'un tel projet, et quels sont les outils et les bibliothèques qu'il faut préparer pour le mener à bien. En ce sens, TANAGRA est plus un outil d'apprentissage des techniques de programmation.

http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html