**Université Lumière Lyon2**
Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

**Pierre-Emmanuel JOUVE**
le 10 décembre 2003

# Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

| | |
|---|---|
| Jean-Paul Rasson, Rapporteur | Professeur, Facultés Universitaires N.D. de la Paix, Namur |
| Gilles Venturini, Rapporteur | Professeur, Université de Tours |
| Mohand-Saïd Hacid, Examinateur | Professeur, Université Claude Bernard-Lyon 1 |
| Michel Lamure, Examinateur | Professeur, Université Claude Bernard-Lyon 1 |
| Gilbert Ritschard, Examinateur | Professeur, Université de Genève |
| Nicolas Nicoloyannis, Directeur de thèse | Professeur, Université Lumière-Lyon 2 |

# Table des matières

# Bibliographie

[AB95] D.W. Aha and R.L. Bankert. A comparative evaluation of sequential feature selection algorithms, 1995.

[AD91] H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, volume 2, pages 547–552, Anaheim, California, 1991. AAAI Press.

[AG92] H. Almuallim and Dietterich T. G. Efficient algorithms for identifying relevant features. Technical Report 92-30-03, 1992.

[AG94] H. Almuallim and Dietterich T. G. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69(1-2):279–305, 1994.

[Arr63] K.J. Arrow. *Social Choice and Individual Values*. 1963.

[Bat94] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks*, 5:537–550, July 1994.

[BBMES03] J. Bi, K. P. Bennett, C. M. Breneman M. Embrechts, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 2003.

[BEF84] J.C. Bezdeck, R. Ehrlich, and W. Full. Fcm:fuzzy c-means algorithm. *Computers and Geoscience*, 1984.

[Ber02] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression trees, The Wadsworth Statistics/Probability Series, Wadsworth, Belmont, CA*. 1984.

[BFOS01] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees. 2001.

[BGG+99] D. Boley, M. Gini, R. Gross, E. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. Partitionning-based clustering for web document categorization. *Decision Support System*, pages 329–341, 1999.

[BLM86] J.P. Barthelemey, B. Leclerc, and B. Monjardet. On the use of ordered sets in problems of comparison and consensus classification. *Journal of Classification*, 3, 1986.

[Bre96a] L. Breiman. Arcing classifiers. *Annals of Statistics*, 1996.

[Bre96b] L. Breiman. Bagging predictors. *Machine Learning*, 1996.

[Car93] C. Cardie. Using decision trees to improve case–based learning. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 25–32. Morgan Kaufmann Publishers, Inc., 1993.

[CCM00] David Cohn, Rich Caruana, and Andrew McCallum. Semi-supervised clustering with user feedback (draft), 2000.

[CF94] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *International Conference on Machine Learning*, pages 28–36, 1994.

[CS02] D. Cristofor and D. Simovici. An information-theoretical approach to clustering categorical databases using genetic algorithms. In *2nd SIAM ICDM, Workshop on clustering high dimensional data*, 2002.

[Dav96] R.N. Dave. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters*, 17:613–623, 1996.

[DB79] D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 1979.

[DBE99] A. Demiriz, K. Bennett, and M. Embrechts. Semi-supervised clustering using genetic algorithms, 1999.

[DCSL02] Manoranjan Dash, Kiseok Choi, Peter Scheuermann, and Huan Liu. Feature selection for clustering - a filter solution. In *Proc. of International Conference on Data Mining (ICDM02)*, pages 115–122, 2002.

[DK82] P. A. Devijver and Kittler. *Pattern Recognition: A Statistical Approach, Englewood Cliffs, New Jersey: Prentice-Hall*. 1982.

[DL97] M. Dash and H. Liu. Feature selection for classification, 1997.

[DL03] Manoranjan Dash and Huan Liu. Feature selection for clustering. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD03)*, pages 110–121, 2003.

[DLP82] Diday, Lemaire, and Pouget. Testu : Eléments d'analyse de données, dunod. 1982.

[DM91] R.L. De Mantaras. A distance-based attribute selection measure for decision tree induction. In *Machine Learning*, volume 6, pages 81–92, 6-9 1991.

[Doa92] J. Doak. An evaluation of feature selection methods and their application to computer security. In Department of Computer Science Davis, CA: University of California, editor, *Technical Report CSE-92-18*, 1992.

[Dom01] B. Dom. An information-theoretic external cluster-validity measure. Technical report, IBM, 2001.

[Dun74] J.C. Dunn. Well separated clusters and optimal fuzzy partitions. *J. Cybern.*, 4:95–104, 1974.

[EC02] Vladimir Estivill-Castro. Why so many clustering algorithms - a position paper. *SIGKDD Explorations*, 4:65–75, 2002.

[EKS+98] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, and Xiaowei Xu. Incremental clustering for mining in a data warehousing environment. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 323–333, 24–27 1998.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[FI92] U. M. Fayyad and K. B. Irani. The attribute selection problem in decision tree generation. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 104–110, 1992.

[Fis87] D. Fisher. Cobweb: Knowledge acquisition via conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[FPSS96] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *A.I. Magazine*, 17:37–54, 1996.

[FPSSU96] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in knowledge discovery and data mining. *AAAI Press*, 1996.

[FRB98] U.M. Fayyad, C. Reina, and P.S. Bradley. Initialization of iterative refinement clustering algorithms. *Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining, KDD98, AAAI Press*, 1998.

[Fre02] A.A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. 2002.

[Fri94] J.H. Friedman. An overview of predictive learning and function approximation. *From Statistics to Neural Networks, Proc. NATO/ASI Workshop, Springer-Verlag*, pages 1–61, 1994.

[FS96] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *Machine Learning : proceedings of the Thirteenth International Conference*, 1996.

[FS97] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *Journal of Computer and System Sciences*, 1997.

[GC85] M. A. Gluck and J. E. Corter. Information, uncertainty, and the utility of categories. *In Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pages 283–287, 1985.

[GD91] K. C. Gowda and Edwin Diday. Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, 24(6):567–578, 1991.

[GE03] Isabelle Guyon and Andre Eliseef Eds. *Journal on Machine Learning Research : Special Issue on Variable and Feature Selection*. 2003.

[Geu03] Pierre Geurts. Traitements de données volumineuses par ensembles d'arbres aléatoires. *Session Spéciale Entreposage et Fouille de Données, XXXVème Journées de la Société Francophone De Statistiques*, pages 111–122, 2003.

[Gha00] B. Ghattas. Agrégation d'arbres de classification. *Revue de Statistique Appliquée*, 2000.

[GKR00] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Clustering categorical data: An approach based on dynamical systems. *VLDB Journal: Very Large Data Bases*, 8(3–4):222–236, 2000.

[Gra89] C. W. J. Granger. Combining forecasts, twenty years later. *Journal of Forecasting*, 1989.

[GRS98] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. CURE: an efficient clustering algorithm for large databases. *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 73–84, 1998.

[GRS00] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.

[GSCWS99] César Guerra-Salcedo, Stephen Chen, Darrell Whitley, and Stephen Smith. Fast and accurate feature selection using hybrid genetic strategies. In Peter J. Angeline, Zbyszek Michalewicz, Marc Schoenauer, Xin Yao, and Ali Zalzala, editors, *Proceedings of the Congress on Evolutionary Computation*, volume 1, pages 177–184, Mayflower Hotel, Washington D.C., USA, 6-9 1999. IEEE Press.

[GSS02] J. Ghosh, A. Strehl, and Merugu S. A consensus framework for integrating distributed clusterings under limited knowledge sharing. *in Proc. of NSF Workshop on Next Generation Data Mining*, 2002.

[GV98] A.D. Gordon and M. Vichi. Partitions of partitions. *Journal of Classification*, (15):265–285, 1998.

[GW89] M. Grötschel and Y. Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming*, 45:59–96, 1989.

[Hal00a] Maria Halkidi. Quality assessment and uncertainty handling in data mining process. In *EDBT PhD Workshop*, 2000.

[Hal00b] Mark A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proc. 17th International Conf. on Machine Learning*, pages 359–366. Morgan Kaufmann, San Francisco, CA, 2000.

[Har84] A. Hart. Experience in the use of an inductive system in knoowledge eng. In M. Bramer, editor, *Research and Development in Expert Systems*. Cambridge Univ. Press, Cambridge, MA,, 1984.

[HBV01] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information System*, 17(2–3):107–145, 2001.

[HBV02a] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part i. *ACM SIGMOD Record*, 31(2):40–45, 2002.

[HBV02b] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering validity checking methods: part ii. *ACM SIGMOD Record*, 31(3):19–27, 2002.

[HK01] J. Han and M. Kamber. Data mining : Concepts and techniques, morgan kaufmann publishers, usa. 2001.

[Hon94] S.J. Hong. Use of contextual information to feature ranking and discretization. In *IEEE Trans. On knowledge and Data Engeneering*, 1994.

[HT01] S. Hirano and Tsumoto. Indiscernibility degrees of objects for evaluating simplicity of knowledge in the clustering procedure. *IEEE International Conference on Data Mining (IEEE ICDM01*, 2001.

[HTO+02] S. Hirano, S. Tsumoto, T. Okuzaki, Y. Hata, and K. Tsumoto. Analysis of biochemical data aided by a rough sets-based clustering technique. *International Journal of Fuzzy Systems*, 4, 2002.

[Hua97] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Research Issues on Data Mining and Knowledge Discovery*, 1997.

[HV01] Maria Halkidi and Michalis Vazirgiannis. Clustering validity assessment : Finding the optimal partioning of a data set. *Proceedings of ICDM01*, pages 187–194, 2001.

[JD88] A.K. Jain and R.C. Dubes. Algorithms for clustering data, englewood cliffs, nj: Prentice hall. 1988.

[JK99] E. Johnson and H. Kargupta. Collective, hierarchical clustering from distributed, heterogeneous data. *Zaki, M., and Ho, C., eds., Large-Scale Parallel KDD Systems, Springer-Verlag LNCS*, 1759, 1999.

[JKP94] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994. Journal version in AIJ, available at http://citeseer.nj.nec.com/13663.html.

[JKSK97] B.H. Jun, C.S. Kim, H.Y. Song, and J. Kim. A new criterion in selection and discretization of attributes for the generation of decision trees. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 19, pages 1371–1375, 1997.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[JN03a] Pierre Emmanuel Jouve and Nicolas Nicoloyannis. Classification non supervisée pour données catégorielles. In *Actes Session Spéciale des XXXVèmes Journées de Statistiques : Entreposage et Fouille de Données*, 2003.

[JN03b] Pierre Emmanuel Jouve and Nicolas Nicoloyannis. Classification non supervisée pour données catégorielles. In *Actes de XXXVèmes Journées de Statistiques*, 2003.

[JN03c] Pierre Emmanuel Jouve and Nicolas Nicoloyannis. Kerouac : an algorithm for clustering categorical data sets with practical advantages. In *Proc. of International Workshop on Data Mining for Actionable Knowledge (PAKDD03)*, 2003.

[JN03d] Pierre Emmanuel Jouve and Nicolas Nicoloyannis. A method for aggregating partitions, applications in knowledge discovery in databases. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD03)*, 2003.

[JN03e] Pierre Emmanuel Jouve and Nicolas Nicoloyannis. A new method for combining partitions, applications for distributed clustering. In *International Workshop on Paralell and Distributed Machine Learning and Data Mining (ECML/PKDD03)*, 2003.

[JN03f] Pierre Emmanuel Jouve and Nicolas Nicoloyannis. The 'who is it ?' problem, application for customizable web sites. In *Proc. of Atlantic Web Intelligence Conference (AWIC'03)*, 2003.

[KARS97] G. Karypis, Aggarwal, V. R., Kumar, and S. Shekhar. Multilevel hypergraph partitioning : Application in vlsi domain. *Proc. of the Design & Automation Conf*, 1997.

[KC00] H. Kargupta and P. editors Chan. Advances in distributed and parallel knowledge discovery. aaai/mit press, cambridge, ma. 2000.

[Ken39] Kendall. A new measure of rank correlation. In *Biometrika N°30*, 1939.

[KL03] Youngok Kim and Soowon Lee. A clustering validity assessment index. *Proceedings of PAKKD03*, pages 602–608, 2003.

[Koh89] T. Kohonen. Self organizing memory. *3rd ed. Springer information sciences series, Springer Verlag, New-York*, 1989.

[Koh96] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page to appear, 1996.

[Kon94] Igor Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning*, pages 171–182, 1994.

[KR92a] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In MIT Press, editor, *Tenth National Conference on Artificial Intelligence*, pages 129–134, 1992.

[KR92b] K. Kira and L.A. Rendell. A practical approach to feature selection. In Morgan Kaufmann, editor, *Proceedings of the Tenth International Conference on Machine Learning*, 1992.

[KR02] J. Kittler and F. editors Roli. *Multiple Classifier Systems*, volume 2634. 2002.

[KS96] D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996.

[KT88] Yves Kodratoff and G. Tecuci. Learning based on conceptual distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):897–909, 1988.

[LD01] Erel Levine and Eytan Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.

[LEB02] Gaëlle Legrand, Walid Erray, and Marc Boullé. Un survey des méthodes de sélection d'attributs dans le data mining. In *Rencontres de la société française de classification, SFC02*, 2002.

[LJF02] M. Law, A.K. Jain, and M. Figueiredo. Feature selection in mixture-based clustering. In *Proc. of Neural Information Processing Systems - NIPS'2002*, 2002.

[LM98] H. Liu and H. Motoda. *Feature Extraction, Construction, and Selection : A Data Mining Perspective, Kluwer Academic, Boston, MA*. 1998.

[LMD98] Huan Liu, Hiroshi Motoda, and Manoranjan Dash. A monotonic measure for optimal feature selection. In *European Conference on Machine Learning*, pages 101–106, 1998.

[LMF02] Nada Lavrac, Hiroshi Motoda, and Tom Fawcett. First international workshop on data mining lessons learned (dmll-2002). *Nineteenth International Conference on Machine Learning (ICML-2002)*, 2002.

[LR00] S. Lallich and R. Rakotomalala. Fast feature selection using partial correlation for multi-valued attributes. In *Proceedings of the 4th European Conference on Knowledge Discovery in Databases, PKDD 2000*, pages 221–231, 2000.

[LS94] Pat Langley and Stephanie Sage. Oblivious decision trees and abstract cases. In *Working Notes of the AAAI-94 Workshop on Case-Based Reasoning*. AAAI Press, 1994.

[LS95] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes, 1995.

[LS96] Huan Liu and Rudy Setiono. A probabilistic approach to feature selection - a filter solution. In *International Conference on Machine Learning*, pages 319–327, 1996.

[Mac67] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability*, volume I : Statistics, pages 281–297, 1967.

[Mar84a] F. Marcotochino. Utilisation des comparaisons par paires en statistique des contingences, etude n°f-071. Technical report, Centre Scientifique IBM-France, Février 1984.

[Mar84b] F. Marcotochino. Utilisation des comparaisons par paires en statistique des contingences, partie ii etude n°f-071. Technical report, Centre Scientifique IBM-France, Mai 1984.

[Mic82] P. Michaud. Agrégation à la majorité 1 : Hommage à condorcet, etude n°f-051. Technical report, Centre Scientifique IBM-France, 1982.

[Mic83] Pierre Michaud. Opinions agregations. *New Trends in Data Analysis and Applications, North-Holland, Amsterdam, J. Janssen, J.F. Marcotorchino and J.M. Proth*, pages 5–27, 1983.

[Mic85] P. Michaud. Agrégation à la majorité 2 : Analyse du résultat d'un vote, etude n°f-051. Technical report, Centre Scientifique IBM-France, 1985.

[Mic87] Pierre Michaud. Condorcet - a man of the avant-garde. *Applied Stochastic Models and Data Analysis, Wiley Chichester, J. Janssen, J.F. Marcotorchino, J.M. Proth and P. Purdue*, 3(3), 1987.

[Mic91] Pierre Michaud. Simulated computation in automatic classification. *Proc. 2nd Symp. on High Performance Computing, M. Durand and F. El Dabaghi*, 1991.

[Mic97] Pierre Michaud. Clustering techniques. *Future Generation Computer Systems*, 13(2–3):135–147, November 1997.

[Min87] J. Mingers. Expert systems – rule induction with statistical data. In *Journal of the Operational Research Society*, 1987.

[Mir01] B. Mirkin. Reinterpreting the category utility function. *Machine Learning*, 42:219–228, 2001.

[ML94] A.W. Moore and M.S. Lee. Efficient algorithms for minimizing cross validation error. In *Proceedings of the 11th Interntional Conf. on Machine Learning*, 1994.

[MM81] F. Marcotorchino and P. Michaud. Heuristic approach to the similarity aggregation problem. *Methods of Operations Research*, 43:395–404, 1981.

[MM96] C. Merz and P. Murphy. Uci repository of machine learning databases. *http://www.ics.uci.edu/#mlearn/mlrepository.html*, 1996.

[Mod93] M. Modrzejewski. Feature selection using rough sets theory. In *Proceedings of the European Conference on Machine Learning*, 1993.

[MS83] R. S. Michalski and R. E. Stepp. Automated construction of classifications : Conceptual clustering versus numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(4):396–410, 1983.

[NF77] P.M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. In *IEEE Transactions Computers*, 1977.

[NH94] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In Jorgeesh Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *20th International Conference on Very Large Data Bases, September 12–15, 1994, Santiago, Chile proceedings*, pages 144–155, Los Altos, CA 94022, USA, 1994. Morgan Kaufmann Publishers.

[Nic88] N. Nicoloyannis. *Structures Prétopologiques et Classification Automatique*. PhD thesis, Université Lyon 1, 1988.

[NTT98] N. Nicoloyannis, M. Terrenoire, and D. Tounissoux. An optimisation model for aggregating preferences : A simulated annealing approach. *Health and System Science*, 2(1-2):33–44, 1998.

[PB97] N.R. Pal and J. Biswas. Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6), 1997.

[PCS00] A. Prodromidis, P. Chan, and S. Stolfo. Meta-learning in distributed data mining systems: Issues and approaches. *H. Kargupta and P. Chan eds, Advances in Distributed & Parallel Knowledge Discovery. AAAI/MIT Press*, 2000.

[PSCF+89] Gregory Piatetsky-Shapiro, Jaime Carbonell, William Frawley, Kamran Parsaye, J. Ross Quinlan, Michael Siegel, and Ramasamy Uthurusamy. Workshop on knowledge discovery in databases. *Fourth International Joint Conference on Artificial Intelligence (IJCAI-1989)*, 1989.

[Qui86] J.R. Quinlan. Introduction of decision trees. In *Machine Learning*, volume 1, pages 81–106, 1986.

[Rak03] A. Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 2003.

[RLR98a] R. Rezaee, B.P.F. Lelieveldt, and J.H.C. Reiber. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters*, 19:237–246, 1998.

[RLR98b] R. Rezaee, B.P.F. Lelieveldt, and J.H.C. Reiber. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters 19*, (237–246), 1998.

[RRJ03] Riquelme J.C. Ruiz R. and Aguilar-Ruiz J.S. Projection-based measure for efficient feature selection. In *Journal of Intelligent and Fuzzy Systems*, 2003.

[SCZ98] C. Sheikholeslami, S. Chatterjee, and A. Zhang. Wavecluster : A multiresolution clustering approach for very large spatial database. *24th VLDB Conference, New York, USA*, 1998.

[SD03] H. Stoppiglia and G. Dreyfus. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research, Special Issue on Variable/Feature Selection*, 2003.

[SG02a] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining partitionings. *Proc. of Conference on Artificial Intelligence (AAAI 2002), Edmonton, AAAI/MIT Press*, 2002.

[SG02b] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR), MIT Press*, 3, 2002.

[Sha48] C.E. Shannon. A mathematical theory of communication. In *Bell System Technical Journal*, 1948.

[Sha96] S.C. Sharma. Applied multivariate techniques, john willey & sons. 1996.

[SJL90] Niblack W. Sheinvald J., Dom B. and Rendell L.A. A modeling approach to feature selection. In *10th International Conf. on Pattern Recognition*, 1990.

[Smy96] P. Smyth. Clustering using monte carlo cross-validation. *Proceedings of KDD, Conference*, 1996.

[Str02] A. Strehl. Relationship-based clustering and cluster ensembles for high-dimensional data mining. *Phd Thesis,University of Texas at Austin*, 2002.

[TK99] S. Theodoridis and K. Koutroubas. Pattern recognition. *Academic Press*, 1999.

[VDJ93] H. Vafaie and K. De Jong. Robust feature selection algorithms. In *Proceedings of the Fifth Conference on Tools for Artificial Intelligence*, pages 356–363, 1993.

[VDJ94] H. Vafaie and H. De Jong. Improving a rule induction system using genetic algorithms. 1994.

[VHD03] Michalis Vazirgiannis, Maria Halkidi, and Gunopulos Dimitrious. *Uncertainty Handling and Quality Assessment*, volume IX of *in Data Mining Series: Advanced Information and Knowledge Processing*. SPRINGER VERLAG, 2003.

[VJ92] H. Vafaie and K. De Jong. Genetic algorithms as a tool for feature selection in machine learning. 1992.

[Wat85] S. Watanabe. Pattern recognition : Human and mechanical. *John Wiley and Sons, Inc., New York*, 1985.

[WL59] W.T. Williams and J.M. Lambert. Multivariate methods in plant ecology. *Journal of Ecology*, 47:83–101, 1959.

[WYM97] Wei Wang, Jiong Yang, and Richard R. Muntz. STING: A statistical information grid approach to spatial data mining. In Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld, editors, *Twenty-Third International Conference on Very Large Data Bases*, pages 186–195, Athens, Greece, 1997. Morgan Kaufmann.

[WYM99] W. Wang, J. Yang, and R Muntz. STING+: An approach to active spatial data mining. In *Fifteenth International Conference on Data Engineering*, pages 116–125, Sydney, Australia, 1999. IEEE Computer Society.

[XB91] X. L Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4), 1991.

[YH98] Jihoon Yang and Vasant Honavar. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, 13:44–49, 1998.

[YPH97] J. Yang, R. Parekh, and V. Honavar. Distal: An inter-pattern distance-based constructive learning algorithm, 1997.

[ZD91] X. Zhou and T.S. Dillon. A statistical–heuristic feature selection criterion for decision tree induction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 13, pages 834–841, 1991.

[Zha64] C.T. Zhan. Approximating symetric relation by equivalence relation. In *SIAM Journal of Applied Mathematics,*, volume 12, 1964.

[ZKV94] D.A. Ziani, Z. Khalil, and R. Vignes. Recherche de sous-ensembles minimaux de variables à partir d'objets symboliques. In *IPMU'94*, volume 2, pages 794–799, 1994.

[ZRL96] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: an efficient data clustering method for very large databases. In *ACM SIGMOD International Conference on Management of Data*, pages 103–114, Montreal, Canada, June 1996.

[ZRR98] A. Zighed, S. Rabaséda, and R. Rakotomalala. Fusinter : a method for discretization of continuous attributes for supervised learning. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 6(33):307–326, 1998.