# Bibliography


[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2–4, 2016.*, pages 265–283, 2016.

[2] Hanady Abdulsalam, David B. Skillicorn, and Patrick Martin. Streaming random forests. In *Eleventh International Database Engineering and Applications Symposium (IDEAS 2007), September 6–8, 2007, Banff, Alberta, Canada*, pages 225–232, 2007.

[3] Marcel R. Ackermann, Marcus Märtens, Christoph Raupach, Kamil Swierkot, Christiane Lammersen, and Christian Sohler. Streamkm++: A clustering algorithm for data streams. *ACM Journal of Experimental Algorithmics*, 17(1), 2012.

[4] Charu C. Aggarwal, editor. *Data Streams – Models and Algorithms*, volume 31 of *Advances in Database Systems*. Springer, 2007.

[5] Charu C. Aggarwal and Jiawei Han, editors. *Frequent Pattern Mining*. Springer, 2014.

[6] Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. A framework for clustering evolving data streams. In *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases, September 9–12, 2003, Berlin, Germany*, pages 81–92, 2003.

[7] Charu C. Aggarwal and Chandan K. Reddy, editors. *Data Clustering: Algorithms and Applications*. CRC Press, 2014.

[8] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Database mining: A performance perspective. *IEEE Trans. Knowl. Data Eng.*, 5(6):914–925, 1993.

[9] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12–15, 1994, Santiago de Chile, Chile*, pages 487–499, 1994.

[10] Tahseen Al-Khateeb, Mohammad M. Masud, Khaled Al-Naami, Sadi Evren Seker, Ahmad M. Mustafa, Latifur Khan, Zouheir Trabelsi, Charu C. Aggarwal, and Jiawei Han. Recurring and novel class detection using class-based ensemble for evolving data stream. *IEEE Trans. Knowl. Data Eng.*, 28(10):2752–2764, 2016.

[11] Khaled Al-Naami, Swarup Chandra, Ahmad M. Mustafa, Latifur Khan, Zhiqiang Lin, Kevin W. Hamlen, and Bhavani M. Thuraisingham. Adaptive encrypted traffic fingerprinting with bi-directional dependence. In *Proceedings of the 32nd Annual Conference on Computer Security Applications, ACSAC 2016, Los Angeles, CA, USA, December 5–9, 2016*, pages 177–188, 2016.

[12] Ezilda Almeida, Carlos Abreu Ferreira, and João Gama. Adaptive model rules from data streams. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23–27, 2013, Proceedings, Part I*, pages 480–492, 2013.

[13] Ezilda Almeida, Carlos Abreu Ferreira, and João Gama. Learning model rules from high-speed data streams. In *Proceedings of the 3rd Workshop on Ubiquitous Data Mining co-located with the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), Beijing, China, August 3, 2013*, page 10, 2013.

[14] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, USA, May 22–24, 1996*, pages 20–29, 1996.

[15] Jaime Andrés-Merino and Lluís Belanche. Streamleader: A new stream clustering algorithm not based in conventional clustering. In *Artificial Neural Networks and Machine*

*Learning, ICANN 2016, 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6–9, 2016, Proceedings, Part II*, pages 208–215, 2016.

[16] Apache Hadoop. `http://hadoop.apache.org`, accessed May 21st, 2017.

[17] Apache Mahout. `http://mahout.apache.org`, accessed May 21st, 2017.

[18] Marta Arias, Albert Bifet, and Alberto Lumbreras. Framework for sentiment analysis of a stream of texts (a 2012 PASCAL Harvest Project). `http://www.cs.upc.edu/~marias/harvest/`, accessed May 28th, 2017.

[19] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7–9, 2007*, pages 1027–1035, 2007.

[20] Ira Assent, Philipp Kranen, Corinna Baldauf, and Thomas Seidl. Anyout: Anytime outlier detection on streaming data. In *Database Systems for Advanced Applications - 17th International Conference, DASFAA 2012, Busan, South Korea, April 15–19, 2012, Proceedings, Part I*, pages 228–242, 2012.

[21] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Web Science 2012, WebSci '12, Evanston, IL, USA, June 22–24, 2012*, pages 33–42, 2012.

[22] Manuel Baena-García, José del Campo-Ávila, Raúl Fidalgo, Albert Bifet, Ricard Gavaldà, and Rafael Morales-Bueno. Early drift detection method. In *Fourth International Workshop on Knowledge Discovery from Data Streams, ECML PKDD*, 2006.

[23] José L. Balcázar, Albert Bifet, and Antoni Lozano. Mining implications from lattices of closed trees. In *Extraction et gestion des connaissances (EGC'2008), Actes des 8èmes journées Extraction et Gestion des Connaissances, Sophia-Antipolis, France, 29 janvier au 1er février 2008, 2 Volumes*, pages 373–384, 2008.

[24] José L. Balcázar, Albert Bifet, and Antoni Lozano. Mining frequent closed rooted trees. *Machine Learning*, 78(1–2):1–33, 2010.

[25] Jean Paul Barddal, Heitor Murilo Gomes, and Fabrício Enembreck. SFNClassifier: A scale-free social network method to handle concept drift. In *Symposium on Applied Computing, SAC 2014, Gyeongju, Republic of Korea, March 24–28, 2014*, pages 786–791, 2014.

[26] Jean Paul Barddal, Heitor Murilo Gomes, and Fabrício Enembreck. SNCStream: A social network-based data stream clustering algorithm. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, April 13–17, 2015*, pages 935–940, 2015.

[27] Michèle Basseville and Igor V. Nikiforov. *Detection of abrupt changes: Theory and application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. `http://people.irisa.fr/Michele.Basseville/kniga/`, accessed May 21st, 2017.

[28] Jürgen Beringer and Eyke Hüllermeier. Efficient instance-based learning on data streams. *Intell. Data Anal.*, 11(6):627–650, 2007.

[29] Daniel Berrar. Confidence curves: An alternative to null hypothesis significance testing for the comparison of classifiers. *Machine Learning*, 106(6):911–949, 2017.

[30] Albert Bifet. *Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams*, volume 207 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2010.

[31] Albert Bifet, Eibe Frank, Geoff Holmes, and Bernhard Pfahringer. Ensembles of restricted Hoeffding trees. *ACM TIST*, 3(2):30:1–30:20, 2012.

[32] Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26–28, 2007, Minneapolis, Minnesota, USA*, pages 443–448, 2007.

[33] Albert Bifet and Ricard Gavaldà. Adaptive learning from evolving data streams. In *Advances in Intelligent Data Analysis VIII, 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31 – September 2, 2009. Proceedings*, pages 249–260, 2009.

[34] Albert Bifet and Ricard Gavaldà. Mining frequent closed trees in evolving data streams. *Intell. Data Anal.*, 15(1):29–48, 2011.

[35] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, and Ricard Gavaldà. Mining frequent closed graphs on evolving data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 11), San Diego, CA, USA, August 21–24, 2011*, pages 591–599, 2011.

[36] Albert Bifet, Geoffrey Holmes, and Bernhard Pfahringer. Leveraging bagging for evolving data streams. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20–24, 2010, Proceedings, Part I*, pages 135–150, 2010.

[37] Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, and Eibe Frank. Fast perceptron decision tree learning from evolving data streams. In *Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21–24, 2010. Proceedings. Part II*, pages 299–310, 2010.

[38] Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, Richard Kirkby, and Ricard Gavaldà. New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 09), Paris, France, June 28 – July 1, 2009*, pages 139–148, 2009.

[39] Albert Bifet, Silviu Maniu, Jianfeng Qian, Guangjian Tian, Cheng He, and Wei Fan. StreamDM: Advanced data mining in Spark streaming. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015, Atlantic City, NJ, USA, November 14–17, 2015*, pages 1608–1611, 2015.

[40] Albert Bifet, Gianmarco De Francisci Morales, Jesse Read, Geoff Holmes, and Bernhard Pfahringer. Efficient online evaluation of big data stream classifiers. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 15), Sydney, NSW, Australia, August 10–13, 2015*, pages 59–68, 2015.

[41] Albert Bifet, Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Indrė Žliobaitė. CD-MOA: Change detection framework for massive online analysis. In *Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013, London, UK, October 17–19, 2013. Proceedings*, pages 92–103, 2013.

[42] Albert Bifet, Jesse Read, Indrė Žliobaitė, Bernhard Pfahringer, and Geoff Holmes. Pitfalls in benchmarking data stream classification and how to avoid them. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23–27, 2013, Proceedings, Part I*, pages 465–479, 2013.

[43] Albert Bifet, Jiajin Zhang, Wei Fan, Cheng He, Jianfeng Zhang, Jianfeng Qian, Geoff Holmes, and Bernhard Pfahringer. Extremely fast decision tree mining for evolving data streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 17), Halifax, Canada, August 14-18, 2017*, 2017, to appear.

[44] Isvani Inocencio Frías Blanco, José del Campo-Ávila, Gonzalo Ramos-Jiménez, Rafael Morales Bueno, Agustín Alejandro Ortiz Díaz, and Yailé Caballero Mota. Online and non-parametric drift detection methods based on Hoeffding's bounds. *IEEE Trans. Knowl. Data Eng.*, 27(3):810–823, 2015.

[45] Joshua Bloch. *Effective Java (2nd Edition) (The Java Series)*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2 edition, 2008.

[46] Christian Bockermann and Hendrik Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012. http://kissen.cs.uni-dortmund.de:8080/PublicPublicationFiles/bockermann_blom_2012c.pdf, accessed May 21st, 2017.

[47] Paolo Boldi, Marco Rosa, and Sebastiano Vigna. HyperANF: Approximating the neighbourhood function of very large graphs on a budget. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 – April 1, 2011*, pages 625–634, 2011.

[48] Marc Boullé. MODL: A Bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1):131–165, 2006.

[49] Christos Boutsidis, Dan Garber, Zohar Shay Karnin, and Edo Liberty. Online principal components analysis. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4–6, 2015*, pages 887–901, 2015.

[50] Robert S. Boyer and J. Strother Moore. MJRTY: A fast majority vote algorithm. In *Automated Reasoning: Essays in Honor of Woody Bledsoe*, pages 105–118, 1991.

[51] Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, and David P. Woodruff. Beating CountSketch for heavy hitters in insertion streams. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18–21, 2016*, pages 740–753, 2016.

[52] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[53] Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–824, 1998.

[54] Leo Breiman. Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36(1–2):85–103, 1999.

[55] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[56] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. The Wadsworth statistics/probability series. Chapman and Hall/CRC, 1984.

[57] Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh, and Edward R. Dougherty. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3):807–824, 2007.

[58] Dariusz Brzezinski and Jerzy Stefanowski. Reacting to different types of concept drift: The accuracy updated ensemble algorithm. *IEEE Trans. Neural Netw. Learning Syst.*, 25(1):81–94, 2014.

[59] Dariusz Brzezinski and Jerzy Stefanowski. Prequential AUC: Properties of the area under the ROC curve for data streams with concept drift. *Knowledge and Information Systems*, 52(2):531–562, 2017.

[60] P. Bühlmann and B. Yu. Analyzing bagging. *Annals of Statistics*, 30:927–961, 2003.

[61] Feng Cao, Martin Ester, Weining Qian, and Aoying Zhou. Density-based clustering over an evolving data stream with noise. In *Proceedings of the Sixth SIAM International Conference on Data Mining, April 20–22, 2006, Bethesda, MD, USA*, pages 328–339, 2006.

[62] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. Apache Flink™: Stream and batch processing in a single engine. *IEEE Data Eng. Bull.*, 38(4):28–38, 2015.

[63] Carlos Castillo. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, New York, NY, USA, 1st edition, 2016.

[64] Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

[65] Shang-Tse Chen, Hsuan-Tien Lin, and Chi-Jen Lu. An online boosting algorithm with theoretical justifications. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 – July 1, 2012*, 2012.

[66] James Cheng, Yiping Ke, and Wilfred Ng. Maintaining frequent closed itemsets over a sliding window. *J. Intell. Inf. Syst.*, 31(3):191–215, 2008.

[67] James Cheng, Yiping Ke, and Wilfred Ng. A survey on algorithms for mining frequent itemsets over data streams. *Knowl. Inf. Syst.*, 16(1):1–27, 2008.

[68] Weiwei Cheng and Eyke Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2–3):211–225, 2009.

[69] Yun Chi, Richard R. Muntz, Siegfried Nijssen, and Joost N. Kok. Frequent subtree mining – an overview. *Fundam. Inform.*, 66(1–2):161–198, 2005.

[70] Yun Chi, Haixun Wang, Philip S. Yu, and Richard R. Muntz. Catch the moment: Maintaining closed frequent itemsets over a data stream sliding window. *Knowl. Inf. Syst.*, 10(3):265–294, 2006.

[71] Kai-Min Chung, Michael Mitzenmacher, and Salil P. Vadhan. Why simple hash functions work: Exploiting the entropy in a data stream. *Theory of Computing*, 9:897–945, 2013.

[72] Amanda Clare and Ross D. King. Knowledge discovery in multi-label phenotype data. In *Principles of Data Mining and Knowledge Discovery, 5th European Conference, PKDD 2001, Freiburg, Germany, September 3–5, 2001, Proceedings*, pages 42–53, 2001.

[73] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The Fast Cauchy Transform and faster robust linear regression. *SIAM J. Comput.*, 45(3):763–810, 2016.

[74] Edith Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. Syst. Sci.*, 55(3):441–453, 1997.

[75] Edith Cohen. All-distances sketches, revisited: HIP estimators for massive graphs analysis. *IEEE Trans. Knowl. Data Eng.*, 27(9):2320–2334, 2015.

[76] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960.

[77] David A. Cohn, Les E. Atlas, and Richard E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[78] Giorgio Corani and Alessio Benavoli. A Bayesian approach for comparing cross-validated algorithms on multiple data sets. *Machine Learning*, 100(2–3):285–304, 2015.

[79] Graham Cormode, Minos N. Garofalakis, Peter J. Haas, and Chris Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, 4(1–3):1–294, 2012.

[80] Graham Cormode and Marios Hadjieleftheriou. Finding the frequent items in streams of data. *Commun. ACM*, 52(10):97–105, 2009.

[81] Graham Cormode and S. Muthu Muthukrishnan. Approximating data with the Count-Min sketch. *IEEE Software*, 29(1):64–69, 2012.

[82] Tamraparni Dasu, Shankar Krishnan, Dongyu Lin, Suresh Venkatasubramanian, and Kevin Yi. Change (detection) you can believe in: Finding distributional shifts in data streams. In *Advances in Intelligent Data Analysis VIII, Proceedings of the 8th International Symposium on Intelligent Data Analysis, IDA 2009*, pages 21–34, 2009.

[83] Mayur Datar, Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Maintaining stream statistics over sliding windows. *SIAM J. Comput.*, 31(6):1794–1813, 2002.

[84] Jonathan de Andrade Silva, Elaine R. Faria, Rodrigo C. Barros, Eduardo R. Hruschka, André Carlos Ponce Leon Ferreira de Carvalho, and João Gama. Data stream clustering: A survey. *ACM Comput. Surv.*, 46(1):13:1–13:31, 2013.

[85] Erico N. de Souza and Stan Matwin. Improvements to Adaboost Dynamic. In *Advances in Artificial Intelligence – 25th Canadian Conference on Artificial Intelligence, Canadian AI 2012, Toronto, ON, Canada, May 28–30, 2012. Proceedings*, pages 293–298, 2012.

[86] Erik D. Demaine, Alejandro López-Ortiz, and J. Ian Munro. Frequency estimation of internet packet streams with limited space. In *Algorithms, ESA 2002, 10th Annual European Symposium, Rome, Italy, September 17–21, 2002, Proceedings*, pages 348–360, 2002.

[87] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res. (JAIR)*, 2:263–286, 1995.

[88] Pedro M. Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 00), Boston, MA, USA, August 20–23, 2000*, pages 71–80, 2000.

[89] Pedro M. Domingos and Geoff Hulten. A general method for scaling up machine learning algorithms and its application to clustering. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 – July 1, 2001*, pages 106–113, 2001.

[90] João Duarte and João Gama. Ensembles of adaptive model rules from high-speed data streams. In *Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine 2014, New York City, USA, August 24, 2014*, pages 198–213, 2014.

[91] João Duarte and João Gama. Multi-target regression from high-speed data streams with adaptive model rules. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19–21, 2015*, pages 1–10, 2015.

[92] João Duarte, João Gama, and Albert Bifet. Adaptive model rules from high-speed data streams. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(3):30:1–30:22, 2016.

[93] Marianne Durand and Philippe Flajolet. Loglog counting of large cardinalities (extended abstract). In *Algorithms, ESA 2003, 11th Annual European Symposium, Budapest, Hungary, September 16–19, 2003, Proceedings*, pages 605–617, 2003.

[94] Benjamin Van Durme and Ashwin Lall. Probabilistic counting with randomized storage. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11–17, 2009*, pages 1574–1579, 2009.

[95] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010.

[96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 96), Portland, Oregon, USA*, pages 226–231, 1996.

[97] Min Fang, Narayanan Shivakumar, Hector Garcia-Molina, Rajeev Motwani, and Jeffrey D. Ullman. Computing iceberg queries efficiently. In *VLDB'98, Proceedings of 24th International Conference on Very Large Data Bases, August 24–27, 1998, New York City, New York, USA*, pages 299–310, 1998.

[98] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 – September 3, 1993*, pages 1022–1029, 1993.

[99] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for *k*-means, PCA and projective clustering. In *Proceedings of the*

*Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1434–1453, 2013.

[100] Alan Fern and Robert Givan. Online ensemble learning: An empirical study. *Machine Learning*, 53(1–2):71–109, 2003.

[101] Hendrik Fichtenberger, Marc Gillé, Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. BICO: BIRCH meets coresets for k-means clustering. In *Algorithms, ESA 2013, 21st Annual European Symposium, Sophia Antipolis, France, September 2–4, 2013. Proceedings*, pages 481–492, 2013.

[102] Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987.

[103] Philippe Flajolet. Approximate counting: A detailed analysis. *BIT*, 25(1):113–134, 1985.

[104] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. In Philippe Jacquet, editor, *2007 Conference on Analysis of Algorithms, AofA07*, Discrete Mathematics and Theoretical Computer Science Proceedings, pages 127–146, 2007.

[105] Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

[106] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, and Hoang Thanh Lam. The SPMF open-source data mining library version 2. In *Machine Learning and Knowledge Discovery in Databases – European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19–23, 2016, Proceedings, Part III*, pages 36–40, 2016.

[107] Eibe Frank, Mark A. Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H. Witten, and Len Trigg. Weka-a machine learning workbench for data mining. In *Data Mining and Knowledge Discovery Handbook, 2nd ed.*, pages 1269–1277. Springer, 2010.

[108] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[109] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[110] João Gama. *Knowledge Discovery from Data Streams*. Chapman and Hall / CRC Data Mining and Knowledge Discovery Series. CRC Press, 2010.

[111] João Gama, Ricardo Fernandes, and Ricardo Rocha. Decision trees for mining data streams. *Intell. Data Anal.*, 10(1):23–45, 2006.

[112] João Gama and Petr Kosina. Recurrent concepts in data streams classification. *Knowl. Inf. Syst.*, 40(3):489–507, 2014.

[113] João Gama and Pedro Medas. Learning decision trees from dynamic data streams. *J. UCS*, 11(8):1353–1366, 2005.

[114] João Gama, Pedro Medas, Gladys Castillo, and Pedro Pereira Rodrigues. Learning with drift detection. In *Advances in Artificial Intelligence – SBIA 2004, 17th Brazilian Symposium on Artificial Intelligence, São Luis, Maranhão, Brazil, September 29 – October 1, 2004, Proceedings*, pages 286–295, 2004.

[115] João Gama, Raquel Sebastião, and Pedro Pereira Rodrigues. Issues in evaluation of stream learning algorithms. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 09), Paris, France, June 28 – July 1, 2009*, pages 329–338, 2009.

[116] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, 2014.

[117] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, December 2012.

[118] Minos N. Garofalakis, Johannes Gehrke, and Rajeev Rastogi, editors. *Data Stream Management – Processing High-Speed Data Streams*. Data-Centric Systems and Applications. Springer, 2016.

[119] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3), 2006.

[120] Dimitrios Georgiadis, Maria Kontaki, Anastasios Gounaris, Apostolos N. Papadopoulos, Kostas Tsichlas, and Yannis Manolopoulos. Continuous outlier detection in data streams: An extensible framework and state-of-the-art algorithms. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22–27, 2013*, pages 1061–1064, 2013.

[121] C. Giannella, J. Han, J. Pei, X. Yan, and P. Yu. Mining frequent patterns in data streams at multiple time granularities. In *Proceedings of the NSF Workshop on Next Generation Data Mining*, pages 191–212, 2002.

[122] Phillip B. Gibbons and Srikanta Tirthapura. Distributed streams algorithms for sliding windows. *Theory of Computing Systems*, 37(3):457–478, 2004.

[123] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26–28, 2004, Proceedings*, pages 22–30, 2004.

[124] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. A survey on ensemble learning for data stream classification. *ACM Comput. Surv.*, 50(2):23:1–36, 2017.

[125] Heitor Murilo Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfharinger, Geoff Holmes, and Talel Abdessalem. Adaptive random forests for evolving data stream classification. *Machine Learning*, 106(9-10):1469–1495, 2017.

[126] Heitor Murilo Gomes and Fabrício Enembreck. SAE2: Advances on the social adaptive ensemble classifier for data streams. In *Symposium on Applied Computing, SAC 2014, Gyeongju, Republic of Korea, March 24–28, 2014*, pages 798–804, 2014.

[127] João Bártolo Gomes, Mohamed Medhat Gaber, Pedro A. C. Sousa, and Ernestina Menasalvas Ruiz. Mining recurring concepts in a dynamic feature space. *IEEE Trans. Neural Netw. Learning Syst.*, 25(1):95–110, 2014.

[128] Vivekanand Gopalkrishnan, David Steier, Harvey Lewis, and James Guszcza. Big data, big business: Bridging the gap. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine 2012). Beijing, China, August 12–12, 2012*, pages 7–11. ACM, 2012.

[129] Michael Greenwald and Sanjeev Khanna. Space-efficient online computation of quantile summaries. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, Santa Barbara, CA, USA, May 21–24, 2001*, pages 58–66, 2001.

[130] Fredrik Gustafsson. *Adaptive Filtering and Change Detection*. Wiley, 2000.

[131] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1):53–87, 2004.

[132] Yang Hang and Simon Fong. Incrementally optimized decision tree for noisy big data. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine 2012, Beijing, China, August 12, 2012*, pages 36–44, 2012.

[133] Ahsanul Haque, Latifur Khan, and Michael Baron. SAND: semi-supervised adaptive novel class detection and classification over data stream. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA.*, pages 1652–1658, 2016.

[134] Stefan Heule, Marc Nunkesser, and Alexander Hall. Hyperloglog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm. In *Joint 2013 EDBT/ICDT Conferences, EDBT '13 Proceedings, Genoa, Italy, March 18–22, 2013*, pages 683–692, 2013.

[135] Geoffrey Holmes, Richard Kirkby, and Bernhard Pfahringer. Stress-testing Hoeffding trees. In *Knowledge Discovery in Databases: PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3–7, 2005, Proceedings*, pages 495–502, 2005.

[136] David Tse Jung Huang, Yun Sing Koh, Gillian Dobbie, and Russel Pears. Detecting volatility shift in data streams. In *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14–17, 2014*, pages 863–868, 2014.

[137] Geoff Hulten and Pedro Domingos. VFML – a toolkit for mining high-speed time-changing data streams. `http://www.cs.washington.edu/dm/vfml/`, accessed May 21st, 2017, 2003.

[138] Geoff Hulten, Laurie Spencer, and Pedro M. Domingos. Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 01), San Francisco, CA, USA, August 26–29, 2001*, pages 97–106, 2001.

[139] Elena Ikonomovska, João Gama, and Saso Dzeroski. Learning model trees from evolving data streams. *Data Min. Knowl. Discov.*, 23(1):128–168, 2011.

[140] Elena Ikonomovska, João Gama, Bernard Zenko, and Saso Dzeroski. Speeding-up Hoeffding-based regression trees with options. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 – July 2, 2011*, pages 537–544, 2011.

[141] Piotr Indyk and David P. Woodruff. Optimal approximations of the frequency moments of data streams. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22–24, 2005*, pages 202–208, 2005.

[142] Chuntao Jiang, Frans Coenen, and Michele Zito. A survey of frequent subgraph mining algorithms. *Knowledge Eng. Review*, 28(1):75–105, 2013.

[143] Paulo Mauricio Gonçalves Jr. and Roberto Souto Maior de Barros. RCD: A recurring concept drift framework. *Pattern Recognition Letters*, 34(9):1018–1025, 2013.

[144] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, June 6–11, 2010, Indianapolis, Indiana, USA*, pages 41–52, 2010.

[145] Richard M. Karp, Scott Shenker, and Christos H. Papadimitriou. A simple algorithm for finding frequent elements in streams and bags. *ACM Trans. Database Syst.*, 28:51–55, 2003.

[146] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[147] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, 1997.

[148] J. Zico Kolter and Marcus A. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research*, 8:2755–2790, 2007.

[149] Nicolas Kourtellis, Gianmarco De Francisci Morales, Albert Bifet, and Arinto Murdopo. VHT: Vertical Hoeffding tree. In *2016 IEEE International Conference on Big Data, Big-Data 2016, Washington DC, USA, December 5–8, 2016*, pages 915–922, 2016.

[150] Philipp Kranen, Ira Assent, Corinna Baldauf, and Thomas Seidl. The ClusTree: Indexing micro-clusters for anytime stream mining. *Knowl. Inf. Syst.*, 29(2):249–272, 2011.

[151] Hardy Kremer, Philipp Kranen, Timm Jansen, Thomas Seidl, Albert Bifet, Geoff Holmes, and Bernhard Pfahringer. An effective evaluation measure for clustering on evolving data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 11), San Diego, CA, USA, August 21–24, 2011*, pages 868–876, 2011.

[152] Jay Kreps. Questioning the lambda architecture, 2014. `https://www.oreilly.com/ideas/questioning-the-lambda-architecture`, accessed May 21st, 2017.

[153] Ludmila I. Kuncheva. Change detection in streaming multivariate data using likelihood detectors. *IEEE Trans. Knowl. Data Eng.*, 25(5):1175–1180, 2013.

[154] Doug Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note, february 2001*, 2001. `https://blogs.gartner.com/doug-laney/`, accessed May 21st, 2017.

[155] Herbert K. H. Lee and Merlise A. Clyde. Lossless online Bayesian bagging. *Journal of Machine Learning Research*, 5:143–151, 2004.

[156] Victor E. Lee, Ruoming Jin, and Gagan Agrawal. Frequent pattern mining in data streams. In *Frequent Pattern Mining*, pages 199–224. Springer, 2014.

[157] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3–6 July 1994 (Special Issue of the SIGIR Forum)*, pages 3–12, 1994.

[158] Hua-Fu Li, Man-Kwan Shan, and Suh-Yin Lee. Online mining of frequent query trees over XML data streams. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23–26, 2006*, pages 959–960, 2006.

[159] Edo Liberty. Simple and deterministic matrix sketching. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013), Chicago, IL, USA, August 11–14, 2013*, pages 581–588, 2013.

[160] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, feb 1994.

[161] Hongyan Liu, Yuan Lin, and Jiawei Han. Methods for mining frequent items in data streams: An overview. *Knowl. Inf. Syst.*, 26(1):1–30, 2011.

[162] Viktor Losing, Barbara Hammer, and Heiko Wersing. KNN classifier with self adjusting memory for heterogeneous concept drift. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12–15, 2016, Barcelona, Spain*, pages 291–300, 2016.

[163] Qiang Ma, S. Muthukrishnan, and Mark Sandler. Frugal streaming for estimating quantiles: One (or two) memory suffices. *CoRR*, abs/1407.1121, 2014.

[164] Nishad Manerikar and Themis Palpanas. Frequent items in streaming data: An experimental evaluation of the state-of-the-art. *Data Knowl. Eng.*, 68(4):415–430, 2009.

[165] Gurmeet Singh Manku and Rajeev Motwani. Approximate frequency counts over data streams. In *VLDB 2002, Proceedings of 28th International Conference on Very Large Data Bases, August 20–23, 2002, Hong Kong, China*, pages 346–357, 2002.

[166] Gurmeet Singh Manku and Rajeev Motwani. Approximate frequency counts over data streams. *Proceedings of the VLDB Endowment*, 5(12):1699, 2012.

[167] Diego Marron, Albert Bifet, and Gianmarco De Francisci Morales. Random forests of very fast decision trees on GPU for mining evolving big data streams. In *ECAI 2014, 21st European Conference on Artificial Intelligence, 18–22 August 2014, Prague, Czech Republic*, pages 615–620, 2014.

[168] Nathan Marz. Storm: distributed and fault-tolerant realtime computation, May 2013. `http://storm-project.net/`, accessed May 21st, 2017.

[169] Nathan Marz and James Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2013.

[170] Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok N. Srivastava, and Nikunj C. Oza. Classification and adaptive novel class detection of feature-evolving data streams. *IEEE Trans. Knowl. Data Eng.*, 25(7):1484–1497, 2013.

[171] Andrew Mccallum and Kamal Nigam. A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[172] Andrew McGregor. Graph stream algorithms: A survey. *SIGMOD Record*, 43(1):9–20, 2014.

[173] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947.

[174] Luiz F. Mendes, Bolin Ding, and Jiawei Han. Stream sequential pattern mining with precise error bounds. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15–19, 2008, Pisa, Italy*, pages 941–946, 2008.

[175] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. Efficient computation of frequent and top-k elements in data streams. In *Database Theory, ICDT 2005, 10th International Conference, Edinburgh, UK, January 5–7, 2005, Proceedings*, pages 398–412, 2005.

[176] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. Why go logarithmic if we can go linear? Towards effective distinct counting of search traffic. In *EDBT 2008, 11th International Conference on Extending Database Technology, Nantes, France, March 25–29, 2008, Proceedings*, pages 618–629, 2008.

[177] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[178] Glenn W. Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.

[179] Jayadev Misra and David Gries. Finding repeated elements. *Sci. Comput. Program.*, 2(2):143–152, 1982.

[180] Carl Mooney and John F. Roddick. Sequential pattern mining – approaches and algorithms. *ACM Comput. Surv.*, 45(2):19:1–19:39, 2013.

[181] Gianmarco De Francisci Morales and Albert Bifet. SAMOA: Scalable Advanced Massive Online Analysis. *Journal of Machine Learning Research*, 16:149–153, 2015.

[182] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaíz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

[183] Robert Morris. Counting large numbers of events in small registers. *Commun. ACM*, 21(10):840–842, 1978.

[184] S. Muthukrishnan, Eric van den Berg, and Yihua Wu. Sequential change detection on data streams. In *Workshops Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28–31, 2007, Omaha, Nebraska, USA*, pages 551–550, 2007.

[185] Hai-Long Nguyen, Yew-Kwong Woon, and Wee Keong Ng. A survey on data stream clustering and classification. *Knowl. Inf. Syst.*, 45(3):535–569, 2015.

[186] Kyosuke Nishida and Koichiro Yamauchi. Detecting concept drift using statistical testing. In *Discovery Science, 10th International Conference, DS 2007, Sendai, Japan, October 1–4, 2007, Proceedings*, pages 264–269, 2007.

[187] David W. Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res. (JAIR)*, 11:169–198, 1999.

[188] Aljaz Osojnik, Pance Panov, and Saso Dzeroski. Multi-label classification via multi-target regression on data streams. *Machine Learning*, 106(6):745–770, 2017.

[189] Nikunj C. Oza and Stuart J. Russell. Experimental comparisons of online and batch versions of bagging and boosting. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 01), San Francisco, CA, USA, August 26–29, 2001*, pages 359–364, 2001.

[190] Nikunj C. Oza and Stuart J. Russell. Online bagging and boosting. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS 2001, Key West, Florida, US, January 4–7, 2001*, 2001.

[191] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[192] Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. ANF: a fast and scalable tool for data mining in massive graphs. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 02), July 23–26, 2002, Edmonton, Alberta, Canada*, pages 81–90, 2002.

[193] Odysseas Papapetrou, Minos N. Garofalakis, and Antonios Deligiannakis. Sketching distributed sliding-window data streams. *The VLDB Journal*, 24(3):345–368, 2015.

[194] Pallabi Parveen, Nate McDaniel, Varun S. Hariharan, Bhavani M. Thuraisingham, and Latifur Khan. Unsupervised ensemble based learning for insider threat detection. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3–5, 2012*, pages 718–727, 2012.

[195] Pallabi Parveen, Nathan McDaniel, Zackary R. Weger, Jonathan Evans, Bhavani M. Thuraisingham, Kevin W. Hamlen, and Latifur Khan. Evolving insider threat detection stream mining perspective. *International Journal on Artificial Intelligence Tools*, 22(5), 2013.

[196] Russel Pears, Sripirakas Sakthithasan, and Yun Sing Koh. Detecting concept change in dynamic data streams – A sequential approach based on reservoir sampling. *Machine Learning*, 97(3):259–293, 2014.

[197] Raphael Pelossof, Michael Jones, Ilia Vovsha, and Cynthia Rudin. Online coordinate boosting. In *On-line Learning for Computer Vision Workshop (OLCV), 2009 IEEE 12th International Conference on Computer Vision*, 2009.

[198] Bernhard Pfahringer, Geoffrey Holmes, and Richard Kirkby. New options for Hoeffding trees. In *AI 2007: Advances in Artificial Intelligence, 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, December 2–6, 2007, Proceedings*, pages 90–99, 2007.

[199] Bernhard Pfahringer, Geoffrey Holmes, and Richard Kirkby. Handling numeric attributes in Hoeffding trees. In *Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008, Osaka, Japan, May 20–23, 2008 Proceedings*, pages 296–307, 2008.

[200] Daryl C. Plummer, Kurt Potter, Richard T. Matlus, Jacqueline Heng, Rolf Jester, Ed Thompson, Adam Sarner, Esteban Kolsky, French Caldwell, John Bace, Neil MacDonald, Brian Gammage, Michael A. Silver, Leslie Fiering, Monica Basso, Ken Dulaney, David Mitchell Smith, Bob Hafner, Mark Fabbi, and Michael A. Bell. Gartner's top predictions for it organizations and users, 2007 and beyond. `https://www.gartner.com/doc/498768/gartners-top-predictions-it-organizations`, 2006.

[201] Abdulhakim Ali Qahtan, Basma Alharbi, Suojin Wang, and Xiangliang Zhang. A PCA-based change detection framework for multidimensional data streams: Change detection in multidimensional data streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 15)*, pages 935–944, 2015.

[202] Massimo Quadrana, Albert Bifet, and Ricard Gavaldà. An efficient closed frequent itemset miner for the MOA stream mining system. *AI Commun.*, 28(1):143–158, 2015.

[203] Arturo Montejo Ráez, Luis Alfonso Ureña López, and Ralf Steinberger. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. In *Advances in Natural Language Processing, 4th International Conference, EsTAL 2004, Alicante, Spain, October 20–22, 2004, Proceedings*, pages 1–12, 2004.

[204] Chedy Raïssi, Pascal Poncelet, and Maguelonne Teisseire. Need for speed : Mining sequential patterns in data streams. In *21èmes Journées Bases de Données Avancées, BDA 2005, Saint Malo, 17–20 octobre 2005, Actes (Informal Proceedings).*, 2005.

[205] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michal Wozniak, and Francisco Herrera. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, 239:39–57, 2017.

[206] T. Ramraj and R. Prabhakar. Frequent subgraph mining algorithms: A survey. *Procedia Computer Science*, 47:197–204, 2015.

[207] Abhik Ray, Larry Holder, and Sutanay Choudhury. Frequent subgraph discovery in large attributed streaming graphs. In *Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine 2014, New York City, USA, August 24, 2014*, pages 166–181, 2014.

[208] Jesse Read, Albert Bifet, Geoff Holmes, and Bernhard Pfahringer. Scalable and efficient multi-label classification for evolving data streams. *Machine Learning*, 88(1–2):243–272, 2012.

[209] Jesse Read, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. Batch-incremental versus instance-incremental learning in dynamic and evolving data. In *Advances in Intelligent Data Analysis XI - 11th International Symposium, IDA 2012, Helsinki, Finland, October 25–27, 2012. Proceedings*, pages 313–323, 2012.

[210] Jesse Read, Bernhard Pfahringer, and Geoffrey Holmes. Multi-label classification using ensembles of pruned sets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15–19, 2008, Pisa, Italy*, pages 995–1000, 2008.

[211] Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7–11, 2009, Proceedings, Part II*, pages 254–269, 2009.

[212] Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. MEKA: A multi-label/multi-target extension to Weka. *Journal of Machine Learning Research*, 17(21):1–5, 2016.

[213] Peter Reutemann and Geoff Holmes. Big data with ADAMS. In *Proceedings of the 4th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine 2015, Sydney, Australia, August 10, 2015*, pages 5–8, 2015.

[214] Peter Reutemann and Joaquin Vanschoren. Scientific workflow management with ADAMS. In *Machine Learning and Knowledge Discovery in Databases – European Conference, ECML PKDD 2012, Bristol, UK, September 24–28, 2012. Proceedings, Part II*, pages 833–837, 2012.

[215] Rocco De Rosa and Nicolò Cesa-Bianchi. Splitting with confidence in decision trees with application to stream mining. In *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12–17, 2015*, pages 1–8, 2015.

[216] Gordon J. Ross, Niall M. Adams, Dimitris K. Tasoulis, and David J. Hand. Exponentially weighted moving average charts for detecting concept drift. *Pattern Recognition Letters*, 33(2):191–198, 2012. Erratum in Pattern Recognition Letters 33:16, 2261 (2012).

[217] Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, and Piotr Duda. A new method for data stream mining based on the misclassification error. *IEEE Trans. Neural Netw. Learning Syst.*, 26(5):1048–1059, 2015.

[218] Leszek Rutkowski, Lena Pietruczuk, Piotr Duda, and Maciej Jaworski. Decision trees for mining data streams based on the McDiarmid's bound. *IEEE Trans. Knowl. Data Eng.*, 25(6):1272–1279, 2013.

[219] Sripirakas Sakthithasan, Russel Pears, Albert Bifet, and Bernhard Pfahringer. Use of ensembles of Fourier spectra in capturing recurrent concepts in data streams. In *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12–17, 2015*, pages 1–8, 2015.

[220] Sripirakas Sakthithasan, Russel Pears, and Yun Sing Koh. One pass concept change detection for data streams. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14–17, 2013, Proceedings, Part II*, pages 461–472, 2013.

[221] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21–24 October 2006, Berkeley, California, USA, Proceedings*, pages 143–152, 2006.

[222] Robert E. Schapire. Using output codes to boost multiclass learning problems. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8–12, 1997*, pages 313–321, 1997.

[223] Jeffrey C. Schlimmer and Richard H. Granger. Incremental learning from noisy data. *Machine Learning*, 1(3):317–354, 1986.

[224] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. `https://research.cs.wisc.edu/techreports/2009/TR1648.pdf`, accessed May 21st, 2017.

[225] Ammar Shaker and Eyke Hüllermeier. IBLStreams: A system for instance-based classification and regression on data streams. *Evolving Systems*, 3(4):235–249, 2012.

[226] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Math. Program.*, 127(1):3–30, 2011.

[227] Jin Shieh and Eamonn J. Keogh. Polishing the right apple: Anytime classification also benefits data streams with constant arrival times. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14–17 December 2010*, pages 461–470, 2010.

[228] R. Smolan and J. Erwitt. *The Human Face of Big Data*. Sterling Publishing Company Incorporated, 2012.

[229] Mohiuddin Solaimani, Mohammed Iftekhar, Latifur Khan, Bhavani M. Thuraisingham, Joey Burton Ingram, and Sadi Evren Seker. Online anomaly detection for multi-source vmware using a distributed streaming framework. *Softw., Pract. Exper.*, 46(11):1479–1497, 2016.

[230] Guojie Song, Dongqing Yang, Bin Cui, Baihua Zheng, Yunfeng Liu, and Kunqing Xie. CLAIM: An efficient method for relaxed frequent closed itemsets mining over stream data. In *Advances in Databases: Concepts, Systems and Applications, 12th International Conference on Database Systems for Advanced Applications, DASFAA 2007, Bangkok, Thailand, April 9–12, 2007, Proceedings*, pages 664–675, 2007.

[231] Mingzhou (Joe) Song and Lin Zhang. Comparison of cluster representations from partial second- to full fourth-order cross moments for data stream clustering. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15–19, 2008, Pisa, Italy*, pages 560–569, 2008.

[232] Ricardo Sousa and João Gama. Online semi-supervised learning for multi-target regression in data streams using AMRules. In *Advances in Intelligent Data Analysis XV – 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13–15, 2016, Proceedings*, pages 123–133, 2016.

[233] W. Nick Street and YongSeog Kim. A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 01), San Francisco, CA, USA, August 26–29, 2001*, pages 377–382, 2001.

[234] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Scalable collaborative filtering approaches for large recommender systems. *Journal of Machine Learning Research*, 10:623–656, 2009.

[235] Grigorios Tsoumakas and Ioannis P. Vlahavas. Random *k*-labelsets: An ensemble method for multilabel classification. In *Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Warsaw, Poland, September 17–21, 2007, Proceedings*, pages 406–417, 2007.

[236] United Nations Global Pulse. Harnessing big data for development and humanitarian action. `http://www.unglobalpulse.org`, accessed May 21st, 2017.

[237] Matthijs van Leeuwen and Arno Siebes. Streamkrimp: Detecting change in data streams. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML/PKDD 2008*, pages 672–687, 2008.

[238] Joaquin Vanschoren, Jan N. van Rijn, and Bernd Bischl. Taking machine learning research online with OpenML. In *Proceedings of the 4th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine 2015, Sydney, Australia, August 10, 2015*, pages 1–4, 2015.

[239] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[240] Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.

[241] Anh Thu Vu, Gianmarco De Francisci Morales, João Gama, and Albert Bifet. Distributed adaptive model rules for mining big data streams. In *2014 IEEE International Conference on Big Data, Big Data 2014, Washington, DC, USA, October 27–30, 2014*, pages 345–353, 2014.

[242] Kiri Wagstaff. Machine learning that matters. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[243] Boyu Wang and Joelle Pineau. Online bagging and boosting for imbalanced data streams. *IEEE Trans. Knowl. Data Eng.*, 28(12):3353–3366, 2016.

[244] Haixun Wang, Wei Fan, Philip S. Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 03), Washington, DC, USA, August 24–27, 2003*, pages 226–235, 2003.

[245] Greg Welch and Gary Bishop. An introduction to the Kalman Filter, Manuscript, 1995. `https://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf`, accessed May 21st, 2017.

[246] Kyu-Young Whang, Brad T. Vander Zanden, and Howard M. Taylor. A linear-time probabilistic counting algorithm for database applications. *ACM Trans. Database Syst.*, 15(2):208–229, 1990.

[247] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[248] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[249] David P. Woodruff. New algorithms for heavy hitters in data streams (invited talk). In *19th International Conference on Database Theory, ICDT 2016, Bordeaux, France, March 15–18, 2016*, pages 4:1–4:12, 2016.

[250] Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD09), Paris, France, June 28 – July 1, 2009*, pages 877–886, 2009.

[251] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9–12 December 2002, Maebashi City, Japan*, pages 721–724, 2002.

[252] Xifeng Yan and Jiawei Han. Closegraph: Mining closed frequent graph patterns. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 03), Washington, DC, USA, August 24–27, 2003*, pages 286–295, 2003.

[253] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache Spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, 2016.

[254] Mohammed Javeed Zaki and Ching-Jiu Hsiao. CHARM: an efficient algorithm for closed itemset mining. In *Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, USA, April 11–13, 2002*, pages 457–473, 2002.

[255] Mohammed Javeed Zaki, Nagender Parimi, Nilanjana De, Feng Gao, Benjarath Phoophakdee, Joe Urban, Vineet Chaoji, Mohammad Al Hasan, and Saeed Salem. Towards generic pattern mining. In *Formal Concept Analysis, Third International Conference, ICFCA 2005, Lens, France, February 14–18, 2005, Proceedings*, pages 1–20, 2005.

[256] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. New algorithms for fast discovery of association rules. In *Proceedings of the Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 97), Newport Beach, California, USA, August 14–17, 1997*, pages 283–286, 1997.

[257] Peng Zhang, Byron J. Gao, Xingquan Zhu, and Li Guo. Enabling fast lazy learning for data streams. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11–14, 2011*, pages 932–941, 2011.

[258] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4–6, 1996.*, pages 103–114, 1996.

[259] Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. Multi-class Adaboost. *Statistics and Its Interface*, 2:349–360, 2009.

[260] Paul Zikopoulos, Chris Eaton, Dirk deRoos, Tom Deutsch, and George Lapis. *IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Companies, Incorporated, 2011.

[261] Indrė Žliobaitė, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. Active learning with evolving streaming data. In *Machine Learning and Knowledge Discovery in Databases,*

*European Conference, ECML PKDD 2011, Athens, Greece, September 5–9, 2011, Proceedings, Part III*, pages 597–612, 2011.

[262] Indrė Žliobaitė, Albert Bifet, Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Evaluation methods and decision theory for classification of streaming data with temporal dependence. *Machine Learning*, 98(3):455–482, 2015.

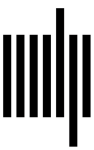
# Machine Learning for Data Streams

## with Practical Examples in MOA

By: Albert Bifet, Ricard Gavaldà, Geoffrey Holmes, Bernhard Pfahringer

## Citation:

*Machine Learning for Data Streams: with Practical Examples in MOA*
By: Albert Bifet, Ricard Gavaldà, Geoffrey Holmes, Bernhard Pfahringer
DOI: 10.7551/mitpress/10654.001.0001
ISBN (electronic): 9780262346047
Publisher: The MIT Press
Published: 2023



The MIT Press

This book was set in Times Roman and Mathtime Pro 2 by the authors.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data is available

ISBN: 978-0-262-03779-2

10 9 8 7 6 5 4 3 2 1