ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Learning and modelling big data

Barbara Hammer[1], Haibo He[2], and Thomas Martinetz[3] *

1- CITEC centre of excellence, Bielefeld University, Germany
2- CISA Lab, University of Rhode Island, USA
3- Institute for Neuro- and Bioinformatics, University of Luebeck, Germany

Abstract. Caused by powerful sensors, advanced digitalisation techniques, and dramatically increased storage capabilities, big data in the sense of large or streaming data sets, very high dimensionality, or complex data formats constitute one of the major challenges faced by machine learning today. In this realm, a couple of typical assumptions of machine learning can no longer be met, such as e.g. the possibility to deal with all data in batch mode or data being identically distributed; this causes the need for novel algorithmic developments and paradigm shifts, or for the adaptation of existing ones to cope with such situations. The goal of this tutorial is to give an overview about recent machine learning approaches for big data, with a focus on principled algorithmic ideas in the field.

1 Introduction

Big data, also referred to as massive data, has been proclaimed as one of the major challenges of the current decade [63, 72, 75, 103, 120]. Recent investigations identify the quantity of data which can be handled in the range of exabytes [61], the article [63] judges the amount of digital data stored worldwide around 13 trillion bytes in 2013. World's technological capacity to store, communicate, and compute information is ever increasing, supported by large storage spaces such as the Utah data centre being built [115, 119]. Research in big data deals with all aspects how to capture, curate, store, search, share, transfer, analyse, and visualise such amounts of data. The problem of big data is not new: areas which traditionally face big data include astronomy, genomics, meteorology, or physical simulations [16, 45, 69, 133]; besides these areas, new domains emerge such as social networks, internet search, finance, or telecommunication; big data carries the promise to substantially enhance decision making e.g. for health care, employment, economic productivity, crime, security, natural disaster, or resource management [103]. At the same time, it opens new challenges e.g. concerning privacy, inter-operability, or general methodology [12, 103, 118]. As an example, the question occurs how to act against possible discrimination caused by grouping or feature relevance determination based on automated big data analytics [5]. Novel algorithmic challenges such as dynamic multi objective optimisation occur [13, 25], and a different philosophy how to handle big data such as demand driven processing only emerges [136].

What exactly is referred to as 'big' depends on the situation at hand, and the term 'big' addresses different quantities: the number of data points, but also

*This research has been supported within the DFG Priority Programme SPP 1527, grant number HA 2716/6-1 to BH and grant number MA 2401/2-1 to TM. Further, HH and BH gratefully acknowledge the support from NSF-DFG Collaborative Research on 'Autonomous Learning', a supplement grant to CNS 1117314.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

its dimensionality or complexity. Douglas Laney primed the characteristic three big V's of big data, nowadays often extended to four V's [83]:

Volume refers to the size of data sets caused by the number of data points, its dimensionality, or both. Volume poses particular challenges on the storage technology and data analysis: access to the data is severely limited, causing the need for online learning and limited memory methods. For distributed sources, sampling can be difficult and strong biases or trends can occur.

Velocity refers to the speed of data accumulation including phenomena such as concept drift, and the need for rapid model adaptation and lifelong learning. Data are often given as streams only, possibly leading to severe non-stationarity or heavy tails of the underlying distribution.

Variety refers to heterogeneous data formats, caused by distributed data sources, highly variable data gathering, different representation technologies, multiple sensors, etc. Machine learning models have to deal with heterogeneous sources, missing values, and different types of data normalisation.

Veracity refers to the fact that data quality can vary significantly for big data sources, and manual curation is usually impossible. Distributed sampling can lead to strong sampling biases or adversarial actions. The problem occurs how to judge this quality and how to deal with quality differences.

From a mathematical perspective, these challenges are mirrored by the following technological issues as detailed in [103]: Dealing with highly distributed data sources; Tracking data provenance, from data generation through data preparation; Validating data; Coping with sampling biases and heterogeneity; Working with different data formats and structures; Developing algorithms that exploit parallel and distributed architectures; Ensuring data integrity; Ensuring data security; Enabling data discovery and integration; Enabling data sharing; Developing methods for visualising massive data; Developing scalable and incremental algorithms; Coping with the need for real-time analysis and decision-making.

In the context of big data, the term 'machine learning' occurs in two roles: *(i) Machine learning as enabling technology:* Machine learning constitutes a major technology to deal with big data. Due to the sheer size of the data, traditional modelling or manual inspection become impossible. Machine learning methods together with their strong mathematical background offer promising possibilities to extract provably accurate information from such data. Hence big data provides conditions such that the use of machine learning techniques instead of more traditional modelling becomes advisable [34]. *(ii) The need to adapt machine learning techniques for big data:* On the other hand, big data forces machine learning research to step out of the classical setting of comparably narrow learning tasks and i.i.d. data which are available prior to training. Out of ten of the most popular machine learning techniques according to the article [135] only few are actually readily applicable for big data; machine learning research has to face the challenges imposed by big data analysis.

Interestingly, it is not clear whether big data actually facilitates or hinders current machine learning techniques: according to classical learning theory, large

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

data sets avoid the risk of overfitting; the need of a good model parameterisation or suitable priors is weakened due to the sheer size of the data. Universal learning models such as Gaussian Processes become independent of the model prior in the limit of large training sets in traditional settings. In big data analytics, however, fundamental statistical assumptions (such as data being i.id.) are not necessarily met. Hence a thorough model validation is necessary and simple models for big data explanation might even need more carefully designed priors [31].

Albeit machine learning and big data are closely interwoven, it seems that entirely novel algorithmic developments to face the challenges of big data are rare in machine learning research in the last years. An internet search for the keywords 'machine learning' and 'big data' does not point to novel scientific developments in the field in the first places. Instead, companies offering data analysis services and funding initiatives connected to big data are listed. While this fact emphasises the great economic and social relevance of the field, it causes the impression that principled technological developments of big data analytics are often limited to an adaptation of existing methods, and a widely accepted canon of big data machine learning approaches is still lacking. One exception is a notable increase of parallel machine learning implementations mostly on the base of MapReduce [125]. In the following, we will provide an overview about recent algorithmic developments which seem promising in the field, without claiming any completeness as concerns relevant research in this rather heterogeneous field.

2 Data representation

Machine learning is coping with the challenge how to represent big data in a way suited for its efficient use for typical machine learning problems. Two different settings have to be addressed: What to do if a huge number of data points is given? What to do if data are very high dimensional?

The first problem, how to deal with a large number of data points efficiently, is often tackled within the algorithmic setting itself. Possible approaches include sampling methods or efficient data representation strategies for matrix data. We will discuss this issue in more detail in section 3. One problem which is independent of the algorithmic setting concerns optimal data compression for storage. Interesting adaptive machine learning techniques have recently been proposed [110]. Further, alternative data descriptions which represent the topological form of the data rather than storing single data points have been investigated [19].

The second problem is how to represent extremely high dimensional data. High dimensionality is often caused by modern sensor technology such as a high spectral resolution or modern fMRI techniques. In the first place, detailed data descriptions caused by a large number of sensors or a detailled sensor resolution carry the promise of a fine grained and highly informative data representation. It is expected that some insights crucially depend on a high level of detail [44, 46]. On the down side, severe problems arise for typical machine learning algorithms from such high dimensional data descriptions: the curse of dimensionality and possible meaninglessness of formalisations for high dimensional data occur [70].

Feature selection constitutes one classical technique to deal with high dimensional data. Quite a few approaches extend feature selection methods to the context of very high data dimensionality, e.g. investigating applications for

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

microarray data or text data [40, 66], addressing their suitability for highly imbalanced data sets [145], or proposing novel feature selection techniques based on geometric principles or stratified sampling [93, 142]. Another relevant aspect concerns the construction of features for high dimensional data such as e.g. random Fourier features [139]. By relying on a suitable regulariser, e.g. the L_1 norm, feature selection can also be incorporated in the machine learning techniques as proposed in the frame of the support feature machine, for example [78]. In a similar way, techniques such as subspace clustering combine the objective of dimensionality reduction with a machine learning objective [132].

One advantage of feature selection is given by their direct interpretability; more general transformations of the data to low dimensions might be more suitable to capture the relevant information for the given task at hand at the price of a possibly reduced interpretability of the representation. Popular classical linear techniques such as independent or principal component analysis have been extended to the setting of high dimensional data [48], or distributed settings [71]. Similarly, efficient extensions of canonical correlation analysis for high dimensional data exist [27]. Based on the observation proved by Johnson and Lindenstrauss [68] that also random projections preserve important information, in particular proximity of data, random projections provide another method of choice to reduce high data dimensionality due to their low computational complexity and often excellent performance. Interestingly, it is possible to accompany the experimental findings by strong theory in which cases random projections can help [37]. The preservation of similarities also constitutes the primary objective of a procrustean approach for dimensionality reduction [52]. Besides linear projections, nonlinear dimensionality reduction techniques enjoy a wide popularity for direct data visualisation or low-dimensional data preprocessing, provided the dimensionality reduction method is equipped with an explicit mapping prescription, see e.g. the recent overview [49, 127]. One particularly prominent dimensionality reduction technique which has been extended to big data sets is offered by deep learning approaches. Deep learning mimics a biologically inspired hierarchical encoding and decoding of information by means of suitable nonlinear maps. Successful applications range from speech recognition to image processing for big data sets [30, 80, 113, 146].

A very promising alternative to feature transformation and dimensionality reduction is offered by sparse coding techniques which take a generative point of view for efficient data representation. Data are approximated by linear combinations of suitable base functions with sparse coding coefficients. For many high dimensional real life data sets e.g. stemming from vision, efficient sparse coding schemes can be found by modern online training schemes such as proposed in [62, 82]. These ideas can also successfully be integrated into particularly efficient sensing of signals in the frame of compressive sampling [35, 67].

Data are not always vectorial in nature, rather a representation by pairwise similarities or dissimilarities becomes increasingly common since such a representation can better be adapted to dedicated data structures, see [106]. The number of data constitutes a strong bottleneck for a matrix representation collecting the pairwise similarities due to its quadratic growth. Hence approximation methods have been proposed to cope with this issue. Low rank approximation such as provided by the Nyström method offer particularly efficient representations

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

which can often be integrated into learning algorithms as linear time technique, see e.g. [50, 51, 152]. Other popular techniques to efficiently deal with large size matrices include random representations, representations by learned functions, or a more general low-rank column-row decomposition of the matrix [11, 92, 65].

In principle, slim representations of high dimensional data aim at a regularisation to get rid of random effects caused by the high dimensionality. Instead of regularising the data representation, promising approaches regularise the machine learning results taking into account the learning context: transfer learning, as an example, can make use of priorly gathered models and puts a strong bias on the new setting, besides the effect of allowing very efficient, possibly even one-shot learning for novel tasks [33, 109, 131].

3 Learning algorithms

Algorithms which are effective for big data sets are severely restricted as concerns their computational complexity and memory use. Different settings can be relevant: on the data side we can face streaming data which arrives over time, distributed data which cannot be put on a single machine, or very large data which fits on a disk but not in the main memory; on the machine side, typical settings range from single machines, multicores, up to parallel clusters. If single machines are used, big data have to be handled de facto as streaming data due to memory restrictions. In this realm, a necessary prerequisite for suitable algorithms is their at most linear run time, their restriction to at most a single pass over the data, and their only constant memory consumption.

This observation has caused a boost of online, streaming, or incremental algorithms for various popular machine learning techniques such as SOM, PCA, inference, variational Bayes, information fusion, learning of mixture models, or kernel density estimation [15, 17, 28, 38, 55, 76, 77, 87, 102]. Online learning and learning from data streams is mathematically well investigated and explicit learning theoretical guarantees can be derived under certain conditions [6, 20, 21, 56, 141]. In practice, the problem of imbalanced data, concept drift and outliers remains a challenge [23, 57, 42, 43, 143]. Online learning can often directly be derived from machine learning cost functions via gradient techniques. Since suitable regularisers such as sparsity constraints are often not smooth, variations such as proximal maps have to be used [147]. In the presence of concept drift, gradient techniques face the problem of varying stability and plasticity of data. Due to this fact many techniques incorporate sufficient statistics of relevant model parameters such as data centroids [26, 50, 104, 105].

Online and streaming algorithms do not permit an adaptation of model complexity based on classical cross-validation technology, such that a further focus of research deals with self tuning models for big data. Promising approaches address clustering schemes and factor models with adaptive model complexity as well as intelligent kernel adaptation strategies [96, 149, 148]. Apart from incremental techniques, scalability for big data can be reached to some extend based on hierarchical schemes. Recent proposals in this context address clustering and data visualisation techniques [3, 39, 41].

Another popular way to reduce the complexity of training addresses subsampling of the data, hence substituting the (infeasible) big data set by a rep-

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

resentative subsample only. Essentially, Nyström techniques as mentioned in section 2 can be linked to sampling landmarks which are used to represent the full data [81]. Classical methods such as Markov chain Monte Carlo or stochastic variational inference demonstrate the enormous power of suitable sampling approaches [108, 144]. However, the main problem of sampling in the presence of big data consists in the fact that it is not easy to obtain representative samples from distributed sources. Approaches to overcome this problem rely on techniques such as snowball sampling or combinations of sampling with suitable transformation techniques, resulting in promising methods e.g. for outlier detection or low dimensional data representation [32, 60, 91, 116].

Another crucial challenge which algorithms have to face is the heterogeneous quality of big data and missing information. For example, one very prominent setting encountered in big data is the fact that usually only a small part of the data is labeled, since labelling often requires substantial human effort. In consequence, many algorithms for big data address unsupervised learning scenarios, semi-supervised models, self-training or one-class classification approaches, or active learning [54, 64, 89, 123, 129, 130, 134]. One very interesting strategy to increase explicit labelling consists in crowd sourcing, such as proposed in [8, 22]. Naturally, such methods raise the challenge how to assure data quality [88].

4 Efficient approximation schemes

Efficient data representation and machine learning models constitute instances of approximation schemes. In this section, we address yet another type of approximation which can be put under the umbrella of algorithmic approximations based on intelligent data structures and algorithm design. Interestingly, the algorithmic complexity of problems can actually become simpler for big data sets such as recently investigated in the context of learning half spaces [29].

What are algorithmic ways which can boost learning for big data? Often, structural observations allow to cast the given optimisation problem into a simpler form. Some approaches rely on symmetries which allow an efficient lift of algorithms such as demonstrated for belief propagation in [2]. Other techniques explore the connection of machine learning tasks to different mathematical formulations such as cross-relations of clustering and non-negative matrix factorisation [140], or the connection of the Rayleigh-Ritz framework to eigenvalue problems in linear discriminant analysis [151]. Further, several relevant problems allow an efficient approximation such as e.g. dynamic programming methods using kernel techniques [55, 58, 137], or summation for some specific kernels [94]. Some approximation methods can be accompanied by explicit bounds. One prominent example are geometric methods for representing data in sparse form relying on so-called core sets [1, 5]. By using the QP formalisation of these geometric problems, an equivalence to popular machine learning problems such as large margin kernel classification and regression can be established, resulting in linear time SVM techniques with a constant number of support vectors [122].

Another important aspect is how to approximate possibly complex models by sparse alternatives, to guarantee fast model evaluation as well as efficient incremental training based on these approximations. Recent approaches propose

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

different ways to prune the model complexity such as the number of prototypes to represent data or the number of support vectors of an SVM [86, 101, 150].

Many efficient algorithms for big data sets rely on advanced data structures which enable a particularly efficient computation of basic algorithmic ingredients. Popular data structures in this realm range from suffix trees for efficient string comparison, to KD-trees, cover trees, and variants for the efficient grouping of geometric data, up to hashing for efficient nearest neighbour search [4]. As an example application, one can consider methods which rely on pairwise proximities such as modern nonlinear data visualisation. Their naive implementation scales quadratically with the number of points. A considerable speedup can be obtained based on intelligent data structures such as KD-trees and a popular approximation method borrowed from particle systems, the Barnes hut approach, resulting in $\mathcal{O}(n \log n)$ visualisation techniques [126].

5 Parallel implementations

Big data is closely interwoven with an increasing availability of parallel and distributed algorithmic realisations and cloud computing. The existing proposals can roughly be distinguished into two main streams: parallel implementations of specific algorithms, mostly relying on the MapReduce framework, or general purpose software libraries which try to bridge the gap between the basic MapReduce framework and the needs of typical machine learning algorithms. Within the latter realm, one can find algorithms platforms for streaming data [10], software libraries for efficient supervised learning [117], or powerful models which allow an easy parallel realisation of typical machine learning inference tasks connected to graphs [90]. Further general approaches for parallel big data analytics include a cloud implementation of data analytics for dynamic data [112], or a MapReduce realisation of popular data mining methods for shared data [24].

More specific parallel implementations address various machine learning models such as robust regression [97], support vector machines [100], extreme learning machines [59, 128], power iteration clustering [138], vector quantisation [53], k-means [7], as well as popular techniques required for machine learning problems such as distributed optimisation [36], online mini batch gradient descent [111], or eigenvalue solvers [73]. Besides, machine learning has also been investigated in its suitability to optimise parallelisation as regards its efficiency [95]. Further, first approaches address issues beyond the mere implementation such as how to incorporate the reliability of hardware components [18].

6 Applications

The developments as detailed above set the ground for an efficient use of machine learning in the context of big data. Their suitability is mirrored by widespread applications in big data analytics in the last years: recent successful approaches range from traditional areas such as astronomy [14, 107, 114], physical simulations [99], traffic analysis [9, 85], or intrusion detection [84, 98], up to social network analysis [47, 74, 79], automated speech and literature processing [102, 121], or the analysis of data from computational neuroscience [124].

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

References

[1] P. K. Agarwal, S. Har-Peled, Kasturi, R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and Computational Geometry, MSRI*, 1–30. Univ. Press, 2005.

[2] B. Ahmadi, K. C. Kersting, M. Mladenov, and S. Natarajan. Exploiting symmetries for scaling loopy belief propagation and relational training. *Machine Learning*, 92(1):91–132, 2013.

[3] C. Alzate and J. A. K. Suykens. Hierarchical kernel spectral clustering. *Neural Networks*, 35:21–30, 2012.

[4] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, Jan. 2008.

[5] F. Angiulli, S. Basta, S. Lodi, and C. Sartori. Distributed strategies for mining outliers in large data sets. *IEEE TKDE*, 25(7):1520–1532, 2013.

[6] P. Auer. Online learning. In C. Sammut and G. I. Webb, eds, *Encycl. Machine Learning*, 736–743. Springer, 2010.

[7] M.-F. Balcan, S. Ehrlich, and Y. Liang. Distributed k-means and k-median clustering on general communication topologies. In *NIPS 26*, 1995–2003. 2013.

[8] L. Barrington, D. Turnbull, and G. Lanckriet. Game-powered machine learning. *Proc. Nat. Acad. Sciences USA*, 109(17):6411–6416, 2012.

[9] A. Bazzan and F. Klügl. Introduction to intelligent systems in traffic and transportation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 25:1–137, 2013.

[10] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive online analysis. *JMLR*, 11:1601–1604, 2010.

[11] A. Bodor, I. Csabai, M. W. Mahoney, and N. Solymosi. RCUR: an R package for CUR matrix decomposition. *BMC Bioinformatics*, 13:103, 2012.

[12] D. Boyd. Privacy and publicity in the context of big data. In *WWW conference*, 2010.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011.

[14] M. B. Brescia, S. Cavuoti, R. D'Abrusco, G. D. Longo, and A. Mercurio. Photometric redshifts for quasars in multi-band surveys. *Astrophysical Journal*, 772(2), 2013.

[15] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. Jordan. Streaming variational Bayes. *NIPS 26*, 1727–1735. 2013.

[16] G. Brumfied. High energy physics: down the petabyte highway. *Nature*, 469:282–283, 2011.

[17] Y. Cao, H. He, and H. Man. SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps. *IEEE TNNLS*, 23(8):1254–1268, 2012.

[18] M. Carbin, S. Misailovic, and M. Rinard. Verifying quantitative reliability for programs that execute on unreliable hardware. *ACM SIGPLAN Notices*, 48(10):33–52, 2013.

[19] G. Carlsson. Topology and data. *Bulletin AMS*, 46(2):255–308, 2009.

[20] N. Cesa-Bianchi, O. Dekel, and O. Shamir. Online learning with switching costs and other adaptive adversaries. *NIPS 26*, 1160–1168. 2013.

[21] N. Cesa-Bianchi, G. Lugosi. *Prediction, Learning and Games*. Cambridge Univ. Press, 2006.

[22] E. C. Chatzilari, S. B. Nikolopoulos, I. Patras, and I. Kompatsiaris. Enhancing computer vision using the collective intelligence of social media. *Studies in CI*, 331:235–271, 2011.

[23] S. Chen and H. He. Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach. *Evolving Systems*, 2(1):35–50, 2011.

[24] Y. Chen, Z. Qiao, S. Davis, H. Jiang, and K.-C. Li. Pipelined multi-GPU MapReduce for big-data processing. *Studies in CI*, 493:231–246, 2013.

[25] I. Comsa, C. B. Grosan, and S. Yang. Dynamics in the multi-objective subset sum: Analysing the behavior of population based algorithms. *Studies in CI*, 490:299–313, 2013.

[26] G. Cormode and S. Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, Apr. 2005.

[27] R. Cruz-Cano and M.-L. Lee. Fast regularized canonical correlation analysis. *Computational Statistics and Data Analysis*, 70:88–100, 2014.

[28] B. Cseke, M. Opper, and G. Sanguinetti. Approximate inference in latent Gaussian-markov models from continuous time observations. *NIPS 26*, 971–979. 2013.

[29] A. Daniely, N. Linial, and S. Shalev-Shwartz. More data speeds up training time in learning halfspaces over sparse vectors. *NIPS 26*, 145–153. 2013.

[30] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng. Large scale distributed deep networks. *NIPS 25*, 1232–1240. 2012.

[31] V. Dhar. Data science and prediction. *Comm. ACM*, 56(12):64–73, 2013.

[32] P. Dhillon, Y. Lu, D. P. Foster, and L. Ungar. New subsampling algorithms for fast least squares regression. *NIPS*, 360–368. 2013.

[33] L. H. Dicker and D. P. Foster. One-shot learning and big data with n=2. *NIPS 26*, 270–278. 2013.

[34] P. Domingos. A few useful things to know about machine learning. *Comm. ACM*, 55(10):78–87, 2012.

[35] D. L. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. on Inf. Theory*, 59(11):7434–7464, 2013.

[36] J. C. Duchi, A. Agarwal, M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Trans.Automat.Contr.*, 57(3):592–606, 2012.

[37] R. J. Durrant and A. Kaban. Random projections as regularizers: Learning a linear discriminant ensemble from fewer observations than dimensions. *ACML 13*, volume 29 of *JMLR Proceedings*, 17–32. 2013.

[38] R. Dutta, A. Morshed, J. Aryal, C. D'Este, and A. Das. Development of an intelligent environmental knowledge system for sustainable agricultural decision support. *Environmental Modelling and Software*, 52:264–272, 2014.

[39] N. Elmqvist, J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Trans.Vis.Comput.Graph.*, 16(3):439–454, 2010.

[40] N. Elssied, O. Ibrahim, and A. Osman. A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7(3):625–638, 2014.

[41] M. Embrechts, C. Gatti, J. Linton, and B. Roysam. Hierarchical clustering for large data sets. *Studies in CI*, 410:197–233, 2013.

[42] J. Feng, H. Xu, S. Mannor, and S. Yan. Online PCA for contaminated data. *NIPS 26*, 764–772. 2013.

[43] J. Feng, H. Xu, and S. Yan. Online robust PCA via stochastic optimization. *NIPS 26*, 404–412. 2013.

[44] G. Forestier, J. Inglada, C. Wemmert, P. Gançarski. Comparison of optical sensors discrimination ability using spectral libraries. *Int. Journal Remote Sensing*, 34(7):2327–2349, 2013.

[45] M. Francis. Future telescope array drives development of exabyte processing. *Ars Technica*, 2012.

[46] G. Gazzola, C.-A. Chou, M. Jeong, and W. Chaovalitwongse. An introduction to the analysis of functional magnetic resonance imaging data. *Fields Institute Comm.*, 63:131–151, 2012.

[47] F. Giannotti, D. Pedreschi, A. Pentland, P. Lukowicz, D. Kossmann, J. Crowley, and D. Helbing. A planetary nervous system for social mining and collective awareness. *Eur.Phys.Journal: Special Topics*, 214(1):49–75, 2012.

[48] S. Giri, M. Bergés, and A. Rowe. Towards automated appliance recognition using an emf sensor in nilm platforms. *Advanced Engineering Informatics*, 27(4):477–485, 2013.

[49] A. Gisbrecht and B. Hammer. Data visualization by nonlinear dimensionality reduction. *WIREs Data Mining and Knowledge Discovery*, accepted.

[50] A. Gisbrecht, B. Mokbel, F.-M. Schleif, X. Zhu, and B. Hammer. Linear time relational prototype based learning. *Int. J. Neural Syst.*, 22(5), 2012.

[51] A. Gittens and M. W. Mahoney. Revisiting the Nystrom method for improved large-scale machine learning. In *ICML (13)* JMLR Proceedings, 567–575.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

[52] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE TPAMI* 35(12):2916–2929, 2013.

[53] M. Grbovic and S. Vucetic. Decentralized estimation using distortion sensitive learning vector quantization. *Pattern Recognition Letters*, 34(9):963–969, 2013.

[54] M.-Z. Guo, C. C. Deng, Y. Liu, and P. Li. Tri-training and MapReduce-based massive data learning. *International Journal of General Systems*, 40(4):355–380, 2011.

[55] H. He. *Self-Adaptive Systems for Machine Intelligence*, Wiley, 2011.

[56] H. He, S. Chen, K. Li, and X. Xu. Incremental learning from stream data. *IEEE TNN*, 22(12):1901–1914, 2011.

[57] H. He, E.A. Garcia. Learning from Imbalanced Data. *IEEE TKDE*, 21(9):1263-1284, 2009.

[58] H He, Z. Ni, and J. Fu. A three-network architecture for on-line learning and optimization based on adaptive dynamic programming. *Neurocomputing* 78(1):3-13, 2012.

[59] Q. He, T. B. Shang, F. Zhuang, and Z. Shi. Parallel extreme learning machine for regression based on MapReduce. *Neurocomputing*, 102:52–58, 2013.

[60] D. Heckathorn. Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49:11–34, 2002.

[61] M. Hilbert and P. Lopez. The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011.

[62] J. Hocke, K. Labusch, E. Barth, and T. Martinetz. Sparse coding and selected applications. *KI*, 26(4):349–355, 2012.

[63] S. Horvath. Aktueller Begriff - big data. Wissenschaftliche Dienste des Deutschen Bundestages, Nov 2013.

[64] G. Huang, S. Song, J. Gupta, and C. Wu. A second order cone programming approach for semi-supervised learning. *Pattern Recognition*, 46(12):3548–3558, 2013.

[65] K. Ishiguro and K. Takeuchi. Extracting essential structure from data. *NTT Technical Review*, 10(11), 2012.

[66] J. Jeyachidra, M. Punithavalli. A study on statistical based feature selection methods for classification of gene microarray dataset. *Journ.Theo.Applied Inf.Tech.*, 53(1):107–114, 2013.

[67] D. Jiang, J. Guo, X. Wu. Low-complexity distributed multi-view video coding for wireless video sensor networks based on compressive sensing theory. *Neurocomp.*, 120:415–421, 2013.

[68] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conf.modern analysis and prob.,Contemporary Mathematics* 26, 189–206. AMS, 1984.

[69] M. Jones, M. Schildhauer, O. Reichman, and S. Bowers. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics 37*, 1:519–544, 2006.

[70] A. Kabán. Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, 22(1):375–385, 2012.

[71] V. Kadaand A. Negi. Computational and space complexity analysis of subxPCA. *Pattern Recognition*, 46(8):2169–2174, 2013.

[72] A. Kahng. Predicting the future of information technology and society [the road ahead]. *IEEE Design and Test of Computers*, 29(6):101–102, 2012.

[73] U. Kang, B. Meeder, E. Papalexakis, and C. Faloutsos. Heigen: Spectral analysis for billion-scale graphs. *IEEE TKDE*, 26(2):350–362, 2014.

[74] M. Kas and B. Suh. Computational framework for generating visual summaries of topical clusters in Twitter streams. *Studies in CI*, 526:173–199, 2014.

[75] T. Khalil. Big data is a big deal. White House, Sep 2012.

[76] D. I. Kim, P. Gopalan, D. Blei, and E. Sudderth. Efficient online inference for bayesian nonparametric relational models. *NIPS 26*, 962–970. 2013.

[77] Y. Kim, K.-A. Toh, A. Teoh, H.-L. Eng, and W.-Y. Yau. An online learning network for biometric scores fusion. *Neurocomputing*, 102:65–77, 2013.

[78] S. Klement, S. Anders, and T. Martinetz. The support feature machine: Classification with the least number of features and application to neuroimaging data. *Neural Computation*, 25(6):1548–1584, 2013.

[79] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proc.Nat.Acad.Sciences USA*, 110(15):5802–5805, 2013.

[80] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. *NIPS 25*, 1106–1114. 2012.

[81] S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the Nyström method. *J. Mach. Learn. Res.*, 13(1):981–1006, Apr. 2012.

[82] K. Labusch, E. Barth, and T. Martinetz. Soft-competitive learning of sparse codes and its application to image reconstruction. *Neurocomputing*, 74(9):1418–1428, 2011.

[83] D. Laney. 3D data management: Controlling data volume, velocity and variety. META Group, February 2001.

[84] P. Laskov, R. Lippmann. Machine learning in adversarial environments. *Mach.Learning*, 81(2):115–119, 2010.

[85] D. Li, W. An, Z. Gong, S. Luo, Y. Xin, X. Du, X. Cui. Research of internet traffic identification scheme based on machine learning algorithms. *Adv.Inf.Sci.Service Sci.*, 4(8):217–228, 2012.

[86] X. Liang, Y. Ma, Y. He, L. Yu, R.-C. Chen, T. Liu, X. Yang, and T.-S. Chen. Fast pruning superfluous support vectors in svms. *Pattern Recognition Letters*, 34(10):1203–1209, 2013.

[87] D. Lin. Online learning of nonparametric mixture models via sequential variational approximation. *NIPS 26*, 395–403. 2013.

[88] Q. Liu, A. Ihler, and M. Steyvers. Scoring workers in crowdsourcing: How many control questions are enough? *NIPS 26*, 1914–1922. 2013.

[89] Y. Liu, C. Ouyang, and B. Yang. Coreference resolution based on probabilistic graphical model for open information extraction. *Int.Journ.Adv.Comp.Tech.*, 4(13):454–461, 2012.

[90] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Distributed graphlab: A framework for machine learning in the cloud. *PVLDB*, 5(8):716–727, 2012.

[91] Y. Lu, P. Dhillon, D. P. Foster, and L. Ungar. Faster ridge regression via the subsampled randomized hadamard transform. *NIPS 26*, 369–377. 2013.

[92] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[93] P. Maji. A rough hypercuboid approach for feature selection in approximation spaces. *IEEE TKDE*, 26(1):16–29, 2014.

[94] S. Maji, A. Berg, J. Malik. Efficient classification for additive kernel SVMs. *IEEE TPAMI*, 35(1):66–77, 2013.

[95] A. Majkowska, D. Zydek, and L. Koszałka. Task allocation in distributed mesh-connected machine learning system: Simplified busy list algorithm with q-learning based queuing. *Advances in Intelligent Systems and Computing*, 226:763–772, 2013.

[96] R. Mall, R. Langone, and J. Suykens. Kernel spectral clustering for big data networks. *Entropy*, 15(5):1567–1586, 2013.

[97] X. Meng and M. W. Mahoney. Robust regression on mapreduce. In *ICML 13* JMLR Proceedings, 888–896.

[98] Y. Meng and L.-F. Kwok. Enhancing false alarm reduction using voted ensemble selection in intrusion detection. *International Journal of CI Systems*, 6(4):626–638, 2013.

[99] A. Mozaffari, M. Gorji-Bandpy, P. Samadian, and S. Noudeh. Analyzing, controlling, and optimizing damavand power plant operating parameters using a synchronous parallel shuffling self-organized pareto strategy and neural network: A survey. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 226(7):848–866, 2012.

[100] D. Nguyen, M. Erdmann, T. Takeyoshi, G. Hattori, K. Matsumoto, and C. Ono. Training multiple support vector machines for personalized web content filters. *IEICE Trans.Inf.Systems*, E96-D(11):2376–2384, 2013.

[101] K. Nikolaidis, T. Mu, and J. Goulermas. Prototype reduction based on direct weighted pruning. *Pattern*

Recognition Letters, 36(1):22–28, 2014.

[102] E. Oja. Machine learning for big data analytics. *Comm.Computer and Inf.Science*, 384:XIII, 2013.

[103] Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council. *Frontiers in Massive Data Analysis*. National Academic Press, 2013.

[104] K. Ouivirach, S. Gharti, and M. Dailey. Incremental behavior modeling and suspicious activity detection. *Pattern Recognition*, 46(3):671–680, 2013.

[105] A. Owen. Data squashing by empirical likelihood. *DAMI*, 7:101–113, 1999.

[106] E. Pekalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, Singapore, December 2005.

[107] M. Pesenson, I. Pesenson, and B. McCollum. The data big bang and the expanding digital universe: High-dimensional, complex and massive data sets in an inflationary epoch. *Advances in Astronomy*, 2010, 2010.

[108] V. Rao and Y. W. Teh. MCMC for continuous-time discrete-state systems. *NIPS 25*, 710–718. 2012.

[109] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. *NIPS 26*, 46–54. 2013.

[110] U. Seiffert. ANNIE-artificial neural network-based image encoder. *Neurocomputing*, 125:229–235, 2014.

[111] S.Shalev-Shwartz, T.Zhang. Accelerated mini-batch stochastic dual coordinate ascent. *NIPS 26*, 378–385. 2013.

[112] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna. Cloud-based software platform for big data analytics in smart grids. *Computing in Science and Engineering*, 15(4):38–47, 2013.

[113] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher networks for large-scale image classification. *NIPS 26*, 163–171. 2013.

[114] P. Škoda. Astroinformatics: Getting new knowledge from the astronomical data avalanche. *Advances in Intelligent Systems and Computing*, 210:15, 2013.

[115] F. Steve. Utah's $1.5 billion cyber-security center under way. *Deseret News*, Jan 2011.

[116] M. Sugiyama and K. Borgwardt. Rapid distance-based outlier detection via sampling. *NIPS 26*, 467–475. 2013.

[117] A. Tacchetti, P. Mallapragada, L. Rosasco, and M. Santoro. Gurls: A least squares library for supervised learning. *JMLR*, 14:3201–3205, 2013.

[118] C.-H. Tai, P. Yu, D.-N. Yang, and M.-S. e. Chen. Structural diversity for resisting community identification in published social networks. *IEEE TKDE*, 26(1):235–252, 2014.

[119] H. Takahashi and Y. Jimbo. Trends in neural engineering. *IEEJ Transactions on Electronics, Information and Systems*, 133(3):544–549, 2013.

[120] T. Tanaka. Big data application technology: An overview. *IEEJ Transactions on Electronics, Information and Systems*, 133(3):550–553, 2013.

[121] A. Thessen, H. Cui, and D. Mozzherin. Applications of natural language processing in biodiversity science. *Advances in Bioinformatics*, 2012, 2012.

[122] I. W. Tsang, J. T. Kwok, P. ming Cheung, and N. Cristianini. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[123] M. Turchi, T. De Bie, and N. Cristianini. An intelligent web agent that autonomously learns how to translate. *Web Intelligence and Agent Systems*, 10(2):165–178, 2012.

[124] N. Turk-Browne. Functional interactions as big data in the human brain. *Science*, 342(6158):580–584, 2013.

[125] J. Dean, S. Ghemawat. MapReduce: Simplified data processing on large clusters. Google Labs, Dec 2004.

[126] L. van der Maaten. Barnes-hut-SNE, *CoRR*, abs/1301.3342, 2013.

[127] L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality reduction: A comparative review, 2008.

[128] M. van Heeswijk, Y. Miche, E. Oja, and A. Lendasse. GPU-accelerated and parallelized elm ensembles for large-scale regression. *Neurocomputing*, 74(16):2430–2437, 2011.

[129] H. Vu, S. Liu, X. Yang, Z. Li, Y. Ren. Identifying microphone from noisy recordings by using representative instance one class-classification approach. *Journ.Networks*, 7(6):908–917, 2012.

[130] R. Wang, S. Kwong, and D. Chen. Inconsistency-based active learning for support vector machines. *Pattern Recognition*, 45(10):3751–3767, 2012.

[131] X. Wang, M. Wang, and W. Li. Scene-specific pedestrian detection for static video surveillance. *IEEE TPAMI*, 36(2):361–374, 2014.

[132] Y.-X. Wang, H. Xu, and C. Leng. Provable subspace clustering: When LRR meets SSC. *NIPS 26*, 64–72. 2013.

[133] P. Webster. Supercomputing the climate: Nasa's big data mission. *CSC World*, 2012.

[134] S. Wu and S. Wang. Information-theoretic outlier detection for large-scale categorical data. *IEEE TKDE*, 25(3):589–602, 2013.

[135] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37, Dec. 2007.

[136] X. B. Wu, X. Zhu, G.-Q. Wu, and W. Ding. Data mining with big data. *IEEE TKDE*, 26(1):97–107, 2014.

[137] X. Xu, C. Lian, L. Zuo, H. He. Kernel-based approximate dynamic programming for real-time online learning control: An experimental study. *IEEE Trans.Contr.Sys.Techn.*, 22(1):146–156, 2014.

[138] W. Yan, U. Brahmakshatriya, Y. Xue, M. Gilder, and B. Wise. P-PIC: Parallel power iteration clustering for big data. *Journal of Parallel and Distributed Computing*, 73(3):352–359, 2013.

[139] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. *NIPS 25*, 485–493. 2012.

[140] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja. Clustering by nonnegative matrix factorization using graph random walk. *NIPS 26*, 1088–1096, 2012.

[141] Z. Yang, H. Zhang, and E. Oja. Online projective nonnegative matrix factorization for large datasets. *ICONIP (3)*, 285–290. Springer, 2012.

[142] Y. Ye, Q. Wu, J. Zhexue Huang, M. Ng, X. Li. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Rec.*, 46(3):769–787, 2013.

[143] Y.-R. Yeh and Y.-C. Wang. A rank-one update method for least squares linear discriminant analysis with concept drift. *Pattern Recognition*, 46(5):1267–1276, 2013.

[144] J. Yin, Q. Ho, and E. Xing. A scalable approach to probabilistic latent space inference of large-scale networks. *NIPS 26*, 422–430. 2013.

[145] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan. Feature selection for high-dimensional imbalanced data. *Neurocomputing*, 105:3–11, 2013.

[146] K. Yu, L. Jia, Y. Chen, and W. Xu. Deep learning: yesterday, today, and tomorrow. *Jisuanji Yanjiu yu Fazhan/Computer Research and Development*, 50(9):1799–1804, 2013.

[147] Y.-L. Yu. On decomposing the proximal map. *NIPS 26*, 91–99. 2013.

[148] J. Zhao and L. Shi. Automated learning of factor analysis with complete and incomplete data. *Computational Statistics and Data Analysis*, 72:205–218, 2014.

[149] S. b. Zhong, D. Chen, Q. Xu, and T. Chen. Optimizing the Gaussian kernel function with the formulated kernel target alignment criterion for two-class pattern classification. *Pattern Recognition*, 46(7):2045–2054, 2013.

[150] F. Zhu, N. Ye, W. Yu, S. Xu, and G. Li. Boundary detection and sample reduction for one-class support vector machines. *Neurocomputing*, 123:166–173, 2014.

[151] L. Zhu and D.-S. Huang. A Rayleigh-Ritz style method for large-scale discriminant analysis. *Pattern Recognition*, 47(4):1698–1708, 2014.

[152] X. Zhu, A. Gisbrecht, F.-M. Schleif, and B. Hammer. Approximation techniques for clustering dissimilarity data. *Neurocomputing*, 90:72–84, 2012.