ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Sparse representation of data

Thomas Villmann[1], Frank-Michael Schleif[2], and Barbara Hammer[2]

1 - University of Applied Sciences Mittweida, Germany
2 - Bielefeld University, Germany

Abstract. The amount of electronic data available today as well as its dimensionality and complexity increases rapidly in many scientific areas including biology, (bio-)chemistry, medicine, physics and its application fields like robotics, bioinformatics or multimedia technologies. Many of these data sets are very complex but have also a simple inherent structure which allows an appropriate sparse representation and modeling of such data with less or no information loss. Advanced methods are needed to extract these inherent but hidden information. Sparsity can be observed at different levels: sparse representation of data points using e.g. dimensionality reduction for efficient data storage, sparse representation of full data sets using e.g. prototypes to achieve compact models for lifelong learning and sparse models of the underlying data structure using sparse encoding techniques. One main goal is to achieve a human-interpretable representation of the essential information. Sparse representations account for the ubiquitous problem that humans have to deal with ever increasing and inherently unlimited information by means of limited resources such as limited time, memory, or perception abilities. Starting with the seminal paper of Olshausen&Field [40] researchers recognized that sparsity can be used as a fundamental principle to arrive at very efficient information processing models for huge and complex data such as observed e.g. in the visual cortex. Nowadays, sparse models include diverse methods such as relevance learning in prototype based representations, sparse coding neural gas, factor analysis methods, latent semantic indexing, sparse Bayesian networks, relevance vector machines and other. This tutorial paper reviews recent developments in the field.

1 Introduction

In standard tasks of every day life such as driving a car, recognizing and talking to people, listening to music, watching movies, having a nice evening in a pub, people are capable of dealing with enormous amounts of information in real time. Usually, they have no big problems to store and to process the relevant information from complex events (such as what is the amount of money my friend owes me after the last visit to a Belgium brewery) and they can easily infer do's and don'ts from previous situations (such as never lent money after 5 glasses of beer). These human capacities are quite remarkable when regarding the comparably slow processing time of single neurons. Further, the large number of neurons and their connectivity enable a huge number of different possible internal representations of information, humans have to deal with. One key ingredient to tackle large amounts of diverse information using only limited

resources and processing time consists in a sparse representation of information or sparse models for information processing. This way, limited resources are sufficient to capture all relevant information, and compressed representations can easily be integrated into further processing or data storage. Sparsity has been observed as a key mechanism in biological systems such as the visual cortex [40, 41], and researchers use various aspects of sparsity to arrive at efficient technical systems for information processing using only limited resources. Thereby, sparsity can aim at a sparse representation of single data points (which can be addressed using e.g. dimensionality reduction), a full data set (which can be represented e.g. using only a small number of prototypes), or a sparse data generation model (like a sparse generative topographic map). The merits of sparsity are not only an in general increased efficiency of the models measured in terms of processing time and a speedup in processing sparse models, but also smaller storage space, better generalization ability due to the compression of inherent noise, and a better interpretability by humans. In general, sparsity can help to infer the underlying information structure present in the data.

In the following contribution, we review recent technological developments in the context of neural models which rely on sparse representations of data.

2 Measures of sparsity

One fundamental question when considering sparsity in technical systems is how sparsity can be measured by means of an evaluation function. On the one hand, such an objective can help to automatically evaluate the degree of sparsity of different models (and probably pick the sparsest one), on the other hand, explicit evaluation functions of sparsity can be used as an objective of learning algorithms to guide the search towards sparse models. As an example, when representing signals by means of an over-complete system of base functions, the coefficients can be constraint using a sparsity measures as done, for example, in [28]

A variety of sparsity measures has been proposed in the literature. Assume that model parameters are given as a vector. Popular measures to evaluate the sparsity of a vector include

- counting the number of zero entries or entries with values smaller than ϵ,
- the (negative) L_p norm where $p \in (0, 1)$,
- entropy measures and variations thereof,
- the Gini index,

and many others. For more complex models which cannot easily be represented by simple vectors, more elaborate measures of sparsity have to be used. Even for vectors, it is not clear which measures of sparsity are appropriate for practical purposes. The contribution [24] formalizes a few natural conditions such as scale invariance or monotonicity with respect to summation of constants and it shows that many popular evaluation criteria for sparsity do not fulfill these properties.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

The Gini index constitutes one exception for which all stated properties can be verified. Interestingly, sparsity can lead to the fact that different metrics are identical such as e.g. solving under-constrained linear equations using the 0- or 1-norm. This is relevant e.g. for applications in blind source separation [33].

3 Sparse representation for better generalization

Often, data are very high dimensional and a direct inspection would yield invalid results due to accumulated noise and the curse of dimensionality. A remedy is possible if data are sparse, as it is often the case in practical applications, since data are in fact generated by an inherently low-dimensional process and massive (probably nonlinear) correlations of the dimensions can be observed. This sparsity can be used to reveal the inherent structure of the data. However, appropriate priors or regularization measures are necessary to uncover the correct ingredients of the data and not only the noise. Prior knowledge about the shape of the models or their parameters is necessary at this place.

There exists a variety of general data projection methods which allow a projection of high dimensional sparse data into low dimensional space where the data structure can be inspected [48]. In low dimensions, data are usually no longer sparse but occupy the space according to the inherent structure. Principal component analysis (PCA) probably constitutes the most prominent linear dimensionality reduction technique; recent developments in this context concern the selection of a set of optimum directions in a non greedy way [8] or improvements of kernel PCA methods for sparse data [3], for example. Often, standard data sets reveal their inherent structure only if nonlinear projection is used. An overview and comparison of recent nonlinear techniques such as locally linear embedding, Isomap, and so forth can be found e.g. in [32]. Many dimensionality reduction mechanisms require preprocessing such as a determination of the topological neighborhood. The work [52] proposes robust methods to uncover this topology for high dimensional sparse and noisy data.

The realization of sparsity by means of low-dimensionality constitutes a principle which can be observed in various models where, by substituting high dimensional sparse data by low-dimensional representations, the generalization ability of the models can be improved. As an example, the approach [30] uses a low dimensional representation of the space by means of encoder networks and integrates this into a reinforcement learning task with good success. In [55] sparse associative networks which mimic properties of the human visual system achieve efficient and robust image recognition. Feature selection and pruning constitute very popular methods to improve the generalization ability of standard models such as feedforward networks, and a variety of techniques has been proposed, e.g. [54, 49, 44]. The work [9] provides a general framework for feature extraction based on partially least squares and the approach [57] manages to underline feature selection with formal theory, showing consistency of selection algorithms under specific conditions. These methods can serve as a canonical background for further models in machine learning.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

4 Sparse coding

Sparse coding tries to represent given data objects in terms of an alternative base system such that the resulting coefficients are sparse. In consequence, the base system must usually be over-complete and the condition of sparsity serves as a regularizing term which makes the problem well posed. Recent developments regarding the theory of over-complete representation can be found in the work [45]. Sparse coding is of course also a widely used concept in signal compression. Recent progress employing sparsity and over completeness of the basis systems has lead to the novel concept of compressed sensing, using sparsity already during the measurement of high dimensional sparse data, detailed in [10].

Blind source separation (BSS) refers to the problem to detect the original base functions from a mixture signal. Thereby, arbitrary base functions are searched for such that the signal can be derived as an arbitrary linear mixture thereof. Since this general problem is ill-posed, the constraint that the base functions are maximum independent is assumed, leading to classical independent component analysis (ICA). This way, ICA can be used to detect a (usually sparse) number of meaningful underlying sources of a given signal. Novel developments in this context have been proposed in various recent publications such as an extension to changing distributions, an application to a priorly unknown number of probably not dominant sources, or an integration of the biologically more reasonable max- instead of the sum-operator [47, 51, 37, 34]. Interestingly, a combination of ICA with topological constraints can lead to a quite robust dictionary of base functions which display biological plausibility [35]. Non-negative matrix factorization refers to the more general problem to decompose a positive matrix corresponding to the signals into two positive forms. Thereby, additional constraints are required such as maximum independence as in BSS or sparsity or smoothness constraints [39, 7].

Sparse coding neural gas as proposed in the approach [28] takes the view of Olshausen&Field. Assuming maximum sparsity, a dictionary of base functions is learned from the data by means of local eigenvectors, superposed by a neural gas inspired partitioning of the data space which is in analogy to the receptive fields of the human visual system. The contribution [29] extends this approach to more general dictionaries. A foundation of this method in terms of the neural gas cost function can be obtained by referring to matrix learning as investigated in [5]. Alternative ways to arrive at base functions are offered by co-occurence analysis [15] or specific data-based optimization [12]. An overview and comparison of different techniques related to non-negative matrix factorization and sparse coding is offered by the contribution [11].

A sparse representation of data provides a valuable property which also allows humans to inspect high dimensional data manually. However, it is not always clear that the respective sparsest solution carries the most interpretable information, more distributed representations probably being closer to natural sources. An exemplary investigation of this claim can be found in the recent work [14].

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

5 Sparse representation in clustering

Clustering partitions a data set into a collection of clusters which cover the data. Often, the clusters can directly be inspected e.g. by their mean vector or representative prototypes, as is the case in popular methods such as k-means, neural gas, or the self-organizing map. Thus, clustering aims at a sparse representation of data because data objects are linked to (sparse) cluster numbers or prototypes; it is an extreme form of sparse coding. As a consequence, many clustering algorithms integrate sparsity measures such as the class entropy, partition entropy, or the like into the training objective.

Clustering of non-vectorial data constitutes one active topic of research. Here, data are characterized by pairwise data dissimilarities instead of a direct vector representation. Since a full distance matrix occupies quadratic space, algorithms usually require at least quadratic time complexity. In case of sparse dissimilarity matrices, a distributed computation such as present e.g. in affinity propagation can take advantage of this sparsity leading to quite efficient computations [16]. Since data are not embedded in a real vector space, a representation by prototypes is no longer straightforward. Exemplar based clustering methods such as affinity propagation or median clustering represent the data set in terms of exemplary data items [16, 21]. For sparse data, however, this choice restricts the possible cluster assignments. Relational clustering as proposed in the approach [20] takes a different view and represents prototypes in terms of virtual combinations of data points. Since, this way, prototypes are no longer sparse itself, an approximation is necessary to allow easy human inspection. The contribution [19] extends these ideas to the generative topographic map, a statistical alternative to the self-organizing map.

A sparse representation of data in terms of prototypes which represent the single clusters gives rise to an efficient universal scheme for incremental or lifelong learning: instead of the already seen data, prototypes serve as a statistics which compresses the relevant information of previous data for future training steps. This approach has been proposed in the context of neural gas and self-organizing maps in the work [1]. Interestingly, it can be extended to relational clustering mechanisms where, due to the restriction to only subparts of the full dissimilarity matrix, a linear time clustering algorithm results because of this reduction to sparse approximations. A similar view is taken in the work [2]: spectral clustering is efficiently performed by a reduction to a subset and a later extension to the full data which is then given by sparse expansions in terms of the known items. Naturally, this scheme is not restricted to clustering but it can be expanded to any algorithms which represent data in terms of sparse quantities which include sufficient information for further training.

6 Sparse models for classification

Classification deals with data to be grouped according to given classes. Prototype based models constitute classical approaches to arrive at a sparse classifi-

cation by attaching class labels to the prototypes. Learning vector quantization (LVQ) is probably one of the most popular algorithms in this area. Interestingly, the sparsity offered by the prototypes can be accompanied with additional mechanisms to enforce sparsity in the data e.g. by relevance learning (GRLVQ) [22]. Recent extensions of GRLVQ learning schemes include the incorporation of adaptive overall structure parameters such as the number of prototypes [26], a more subtle adaptation of the metric if known structural components are present [59], or more sophisticated choices of the metric to better adapt to the respective application scenario at hand [38].

Naturally, sparsity constraints can also be integrated into alternative classifiers such as simple linear classification schemes. In this context, sparsity can serve as vehicle to efficiently extend the classification schemes to very large data sets, such as proposed in the approach [6] for linear classifiers by means of an appropriate approximation of the likelihood function.

7 Sparse regression models

Regression models extend classification such that arbitrary real vectors can be used as output instead of discrete numbers only. A majority of classical machine learning models has been proposed for general regression tasks including feedforward networks, radial basis function networks, support vector machines, and the like. Since Vapnik's seminal work on structural risk minimization, regularization is included in popular regression models and often naturally yields sparse models. The support vector machine (SVM) which expands the solution in terms of the support vectors, constitutes a prime example. However, with increasing data set size, this number usually increases such that a representation of the solution by a priorly limited number of support vectors constitutes an active area of research in the context of sparsity [23]. The choice of the loss function can serve as an alternative vehicle to arrive at sparser solutions for SVM [56]. Another bottleneck of SVM training and kernel machines in general is the design and computation of the kernel. Approximate computation of the Gram matrix can be based on popular methods such as the Nyström approximation, while several methods for optimum sparse kernel design have recently been presented [18, 46] including optimization of kernel parameters using LASSO, for example.

Unlike prototype based methods, support vectors do not provide representative examples of the data set, rather elements at the boundaries are picked. The relevance vector machine (RVM) combines the benefits of the SVM such as an excellent generalization ability because of margin optimization with an intuitive representation of data in terms of representative examples. This can be further extended such that relevance vectors can lie at general positions in the feature space [17].

Interestingly, it is possible to accompany sparse models with theoretical validity guarantees such as presented in the approach [58] for regularized least squares regression under noise.

Dedicated *numerical* procedures have been proposed in the context of sparse

regression models. These concern different aspects such as an efficient optimization of the overall objective under sparsity constraints [13], an efficient computation of the Gram matrix for data streams or sparse feature spaces [42], efficient training of sparse kernel machines based on attractor dynamics, or general methods to induce sparsity into popular regression models [31] . Altogether, a quite powerful repertoire of state-of-the-art technology is available to infer sparse models from the data.

8 Conclusions

As reviewed in this overview paper, sparsity plays an essential role in quite diverse areas of machine learning, ranging from data representation to clustering, classification, and regression tasks. In these areas, powerful tools have been designed which are partially accompanied by interesting mathematical guarantees on the one hand, and successful real life applications, on the other hand. This includes classification and regression for images based on sparse representation [53], microarray data analysis using sparse component analysis [43], an inference of cognitively relevant visual neighborhoods of objects by sparse coding [36], blind source separation for audio data [25], applications to SAR images [50], or handwritten digit recognition [27], or the inference of extremely sparse time series prediction models [4], to name just a few. However, when enforcing sparsity, care has to be taken such that no relevant information is lost during the process (sparsity in a pub should not lead to sparse beer, for example).

References

[1] N. Alex, A. Hasenfuss, and B. Hammer. Patch clustering for massive data sets. *Neurocomputing*, 72(7-9):1455–1469, 2009.

[2] C. Alzate and J. Suykens. Highly sparse kernel spectral clustering with predictive out-of-sample extensions. In *this volume*.

[3] C. Alzate and J. Suykens. Kernel component analysis using an epsilon-insensitive robust loss function. *IEEE Transactions on Neural Networks*, 19(9):1583–1598, 2008.

[4] C. Archambeau, M. Biga Diambeidou, S. Dablemont, G. de Lannoy, N. Delannay, N. Donckers, D. Francois, C. Krier, J. Lee, G. Simon, and F. Vrins. Blind prediction of ESANN time series, 2006. Talk at ESANN'06 accompanying a reviewed and conditionally accepted contribution (with very hard conditions which could hardly be fulfilled).

[5] B. Arnonkijpanich, A. Hasenfuss, and B. Hammer. Local matrix adaptation in clustering and applications for manifold visualization. *Neural Networks*, to appear.

[6] D. Balakrishnan, S. andf Madigan. Algorithms for sparse linear classifiers in the massive data setting. *Journal of Machine Learning Research*, 9:313–337, 2008.

[7] A. Chichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative matrix and tensor factorization*. Wiley, 2009.

[8] A. D'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

[9] C. Dhanjal, S. Gunn, and J. Shawe-Taylor. Efficient sparse kernel feature extraction based on partial least squares. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1347–1361, 2009.

[10] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289 – 1306, 2006.

[11] D. Dornbusch, R. Haschke, S. Menzel, and H. Wersing. Finding correlations in multimodal data using decomposition approaches. In *this volume*.

[12] Duarte-Carvajalino and G. J.M., Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, 18(7):1395–1408, 2009.

[13] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

[14] M. Elad and I. Yavneh. A plurality of sparse representations is better than the sparsest one alone. *IEEE Transactions on Information Theory*, 55(10):4701–4714, 2009.

[15] J. B. Estrach and S. Mallat. Geometric models with co-occurrence groups. In *this volume*.

[16] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[17] J. Gao, J. Zhang, and D. Tien. Relevance units latent variable model and nonlinear dimensionality reduction. *IEEE Transactions on Neural Networks*, 21(1):123–135, 2010.

[18] P. Gao, J.and Kwan and D. Shi. Sparse kernel learning with lasso and bayesian inference algorithm. *Neural Networks*, 23(2):257–264, 2010.

[19] A. Gisbrecht, B. Mokbel, and B. Hammer. Relational generative topographic mapping. In *this volume*.

[20] B. Hammer and A. Hasenfuss. Clustering very large dissimilarity data sets. In N. El Gayar and F. Schwenker, editors, *ANNPR'2010*, volume 5998 of *Lecture Notes in Artificial Intelligence*, pages 259–273. Springer, 2010.

[21] B. Hammer, A. Hasenfuss, and F. Rossi. Median topographic maps for biological data sets. In M. Biehl, B. Hammer, M. Verleysen, and T. Villmann, editors, *Similarity Based Clustering*, Lecture Notes Artificial Intelligence Vol. 5400, pages 92–117. Springer, 2009.

[22] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[23] M. Hu, Y. Chen, and J.-Y. Kwok. Building sparse multiple-kernel svm classifiers. *IEEE Transactions on Neural Networks*, 20(5):827–839, 2009.

[24] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[25] M. Jafari, E. Vincent, S. Abdallah, M. Plumbley, and M. Davies. An adaptive stereo basis method for convolutive blind audio source separation. *Neurocomputing*, 71(10-12):2087–2097, 2008.

[26] T. Kierzmann, S. Lange, and M. Riedmiller. Incremental grlvq: Learning relevant features for 3d object recognition. *Neurocomputing*, 71(13-15):2868–2879, 2008.

[27] K. Labusch, E. Barth, and T. Martinetz. Simple method for high-performance digit recognition based on sparse coding. *IEEE Transactions on Neural Networks*, 19(11):1985–1989, 2008.

[28] K. Labusch, E. Barth, and T. Martinetz. Sparse coding neural gas: Learning of overcomplete data representations. *Neurocomputing*, 72(7-9):1547–1555, 2009.

[29] K. Labusch and T. Martinetz. Learning sparse codes for image reconstruction. In *this volume*.

[30] S. Lange and M. Riedmiller. Deep learning visual control policies. In *this volume*.

[31] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

[32] J. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, 2007.

[33] Y. Li, A. Cichocki, S.-I. Amari, S. Xie, and C. Guan. Equivalence probability and sparsity of two sparse solutions in sparse representation. *IEEE Transactions on Neural Networks*, 19(12):2009–2021, 2008.

[34] J. Lücke and M. Sahani. Maximal causes for non-linear component extraction. *Journal of Machine Learning Research*, 9:1227–1267, 2008.

[35] L. Ma and L. Zhang. Overcomplete topographic independent component analysis. *Neurocomputing*, 71(10-12):2217–2223, 2008.

[36] J. Miao, L. Duan, L. Qing, W. Gao, X. Chen, and Y. Yuan. Spatial relationship representation for visual object searching. *Neurocomputing*, 71(10-12):1813–1823, 2008.

[37] F. Movahedi Naini, G. Hosein Mohimani, M. Babaie-Zadeh, and C. Jutten. Estimating the mixing matrix in sparse component analysis (sca) based on partial k-dimensional subspace clustering. *Neurocomputing*, 71(10-12):2330–2343, 2008.

[38] E. Mwebaze, P. Schneider, F.-M. Schleif, S. Haase, and T. Villmann. Divergence based learning vector quantization. In *this volume*.

[39] P. O'Grady and B. Pearlmutter. Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. *Neurocomputing*, 72(1-3):88–101, 2008.

[40] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[41] R. Pashaie and N. Farhat. Self-organization in a parametrically coupled logistic map network: A model for information processing in the visual cortex. *IEEE Transactions on Neural Networks*, 20(4):597–608, 2009.

[42] Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, and S. Vishwanathan. Hash kernels for structured data. *Journal of Machine Learning Research*, 10:2615–2637, 2009.

[43] K. Stadlthanner, F. Theis, E. Lang, A. Tome, C. Puntonet, and J. Gorriz. Hybridizing sparse component analysis with genetic algorithms for microarray analysis. *Neurocomputing*, 71(10-12):2356–2376, 2008.

[44] J. Tikka and J. Hollmen. Sequential input selection algorithm for long-term prediction of time series. *Neurocomputing*, 71(13-15):2604–2615, 2008.

[45] P. Tseng. Further results on stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 55(2):888–899, 2009.

[46] D. Tzikas, A. Likas, and N. Galatsanos. Sparse bayesian modeling with adaptive kernel learning. *IEEE Transactions on Neural Networks*, 20(6):926–937, 2009.

[47] D. Tzikas, A. Likas, and N. Galatsanos. Variational bayesian sparse kernel-based blind image deconvolution with student's-t priors. *IEEE Transactions on Image Processing*, 18(4):753–764, 2009.

[48] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[49] M. van Gerven, A. Bahramisharif, T. Heskes, and O. Jensen. Selecting features for bci control based on a covert spatial attention paradigm. *Neural Networks*, 22(9):1271–1277, 2009.

[50] Z.-M. Wang and W.-W. Wang. Fast and adaptive method for sar superresolution imaging based on point scattering model and optimal basis selection. *IEEE Transactions on Image Processing*, 18(7):1477–1486, 2009.

[51] Y. Washizawa and A. Cichocki. Sparse blind identification and separation by using adaptive k-orthodrome clustering. *Neurocomputing*, 71(10-12):2321–2329, 2008.

[52] G. Wen. Relative transformation-based neighborhood optimization for isometric embedding. *Neurocomputing*, 72(4-6):1205–1213, 2009.

[53] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[54] J.-B. Yang, K.-Q. Shen, C.-J. Ong, and X.-P. Li. Feature selection for mlp neural network: The use of random permutation of probabilistic outputs. *IEEE Transactions on Neural Networks*, 20(12):1911–1922, 2009.

[55] X. Zeng, S. Luo, and Q. Li. An associative sparse coding neural network and applications. *Neurocomputing*, 73(4-6):684–689, 2010.

[56] L. Zhang and W. Zhou. On the sparseness of 1-norm support vector machines. *Neural Networks*.

[57] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

[58] S. Zhou, J. Lafferty, and L. Wasserman. Compressed and privacy-sensitive sparse regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009.

[59] D. Zühlke, F.-M. Schleif, T. Geweniger, S. Haase, and T. Villmann. Learning vector quantization for heterogeneous structured data. In *this volume*.