ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Feature and kernel learning

Verónica Bolón-Canedo[1], Michele Donini[2] and Fabio Aiolli[2]

1- University of A Coruña - Department of Computer Science
Campus de Elviña s/n 15071 A Coruña - Spain

2- University of Padova - Department of Mathematics
Via Trieste, 63, 35121 Padova - Italy

Abstract. Feature selection and weighting has been an active research area in the last few decades finding success in many different applications. With the advent of Big Data, the adequate identification of the relevant features has converted feature selection in an even more indispensable step. On the other side, in kernel methods features are implicitly represented by means of feature mappings and kernels. It has been shown that the correct selection of the kernel is a crucial task, as long as an erroneous selection can lead to poor performance. Unfortunately, manually searching for an optimal kernel is a time-consuming and a sub-optimal choice. This tutorial is concerned with the use of data to learn features and kernels automatically. We provide a survey of recent methods developed for feature selection/learning and their application to real world problems, together with a review of the contributions to the ESANN 2015 special session on *Feature and Kernel Learning.*

1 Feature learning

In the last few years, several datasets with high dimensionality have become publicly available on the Internet. This fact has brought an interesting challenge to the research community, since for the machine learning methods it is difficult to deal with a high number of input features. To cope with the problem of the high number of input features, dimensionality reduction techniques can be applied to reduce the dimensionality of the original data and improve learning performance. These dimensionality reduction techniques usually come in two flavors: *feature selection* and *feature extraction.*

Feature selection and feature extraction each have their own merits [1]. On the one hand, feature extraction techniques achieve dimensionality reduction by combining the original features. In this manner, they are able to generate a set of new features, which is usually more compact and of stronger discriminating power. It is preferable in applications such as image analysis, signal processing, and information retrieval, where model accuracy is more important than model interpretability. On the other hand, feature selection achieves dimensionality reduction by removing the irrelevant and redundant features. It is widely used in data mining applications, such as text mining, genetics analysis, and sensor data processing. Due to the fact that feature selection maintains the original features, it is especially useful for applications where the original features are important for model interpretation and knowledge extraction.

Feature selection methods are usually divided into three major approaches based upon the relationship between a feature selection algorithm and the inductive learning method used to infer a model [2]. *Filters* rely on the general characteristics of training data and carry out the feature selection process as a pre-processing step independently from the induction algorithm. On the contrary, *wrappers* involve optimizing a predictor as a part of the selection process. In between them one can find *embedded* methods, which perform feature selection in the process of training and are usually specific to given learning machines. Popular and widely-used feature selection methods are Correlation-based Feature Selection (CFS) [3], minimum Redundancy Maximum Relevance (mRMR) [4], or Support Vector Machine - Recursive Feature Elimination (SVM-RFE) [5], among others.

As for feature extraction, the most popular method is called Principal Component Analysis (PCA) [6], which converts a set of observations of possibly correlated features into a set of values of linearly uncorrelated features called principal components. The number of principal components is less than or equal to the number of original features.

1.1 Recent contributions

There exist numerous papers and books proving the benefits of the feature selection process [2]. In [7], classical feature selection techniques are provided in the form of a basic taxonomy and their applicability to bioinformatics applications is discussed. Another work on comparing state-of-the-art feature selection methods when dealing with thousands of features, using both synthetic data and real data, is presented in [8]. Brown et al. [9] presented a unifying framework for feature selection based on information theory, covering up to 17 different methods. More recently, the performance of well-known feature selection methods in the presence of several complications (such as noise, redundancy or interaction between attributes) was tested in [10].

However, since none of the existing methods mentioned has demonstrated significantly superiority over the others, researchers are usually focused on finding a good method for a specific problem setting. Therefore, new and novel feature selection methods are constantly appearing using different strategies. In the last few years, the review of the literature has shown a tendency to mix algorithms, either in the form of hybrid methods [11, 12, 13, 14] or ensemble methods [15, 16, 17, 18, 19].

1.2 Applications

Feature selection methods are currently being applied to problems of very different areas. In the next paragraphs we will describe some of the most popular applications that are promoting the use of feature selection:

- **Computational biology.** Bioinformatic tools have been widely applied to genomics, proteomics, gene networks, structure prediction, disease diagnosis and drug design. DNA microarrays have been widely used in simultaneously monitoring mRNA expressions of thousands of genes in many areas of biomedical research. These data sets typically consist of several hundred samples as opposed to thousands of genes, whereby feature selection is paramount. Because of this, a myriad of works in the feature selection field have been devoted to help in the classification of DNA microarrays. A complete review of up-to-date feature selection methods developed for dealing with microarray data can be found in [20].

- **Face recognition**. The recognition of a human face has a wide range of applications, such as face-based video indexing and browsing engines, biometric identity authentication, human-computer interaction, and multimedia monitoring/surveillance. An important issue in this field is to determine which features from an image are the most informative for recognition purposes, so feature selection algorithms for face recognition have been recently suggested [21, 22, 23, 24, 25, 26]

- **Health studies**. The recent explosion in data available for analysis is as evident in health care as anywhere else. Private and public insurers, health care providers, particularly hospitals, physician groups and laboratories, and government agencies are able to generate far more digital information than ever before. Many health studies are longitudinal: each subject is followed over a period of time and many covariates and responses of each subject are collected at different time points. Feature selection has proven effective in helping with the diagnosis of several diseases, such as retinopathy of prematurity [27], evaporative dry eye [28], pulmonary nodules [29] or cardiac pacemaker implantation [30], among others.

- **Financial engineering and risk management**. Technological revolution and trade globalization have introduced a new era of financial markets. Over the last three decades, an enormous number of new financial products have been created to meet customer demands. The stock market trend is very complex and is influenced by various factors. Therefore it is very necessary to find out the most significant factors of the stock market and feature selection can be applied to achieve this goal [31, 32, 33, 34].

- **Text classification**. The categorization of documents into a fixed number of predefined categories has become a popular problem in Internet applications such as Spam email or shopping. Each unique word in a document is considered as a feature, so it is highly important to select a subset of all the possible features, allowing to reduce the computational requirements of learning algorithms. In the last few years, a number of works which promote the use of feature selection for text categorization have been presented [35, 36, 37, 38, 11].

2 Kernel learning

Kernel machines and kernel-based algorithms are very popular in machine learning and have shown their state-of-the-art performance. Kernel methods are used to tackle a variety of learning tasks (e.g. classification, regression, clustering and more). In these algorithms the features are provided intrinsically using a positive semi-definite kernel function that can be interpreted as a similarity measure (i.e. a scalar product) in a high dimensional Hilbert Space. The goal of Kernel Learning (KL) and Multiple Kernel Learning (MKL) is to create a machine able to provide automatically good kernels for a particular problem and avoiding, in this way, the arbitrary user's choice of a kernel function.

KL is often adopted within a semi-supervised learning setting [39, 40] and tries to learn the kernel matrix using all the available data (labeled and unlabeled examples) optimizing an objective function that improves the accordance between the kernel and the set of i.i.d. labeled data. Some possible ways that have been investigated to obtain this result have been either maximizing the alignment [41, 42] or exploiting bounds on the *(Local) Rademacher complexity* [43]. Conversely, unlabeled data are typically used to regularize the models avoiding the non-smoothness of the discriminant function.

In the next sections we will review the four principal families of the KL algorithms: parametric methods, transduction, feature extraction in feature space, semi-supervised spectral kernel learning, and multiple kernel learning.

2.1 Parametric methods for kernel learning

This family of algorithms tries to optimize the parameters of a specific kernel function (e.g. RBF, polynomial). For example, in [44] the spread parameter of the RBF kernel is optimized with the *Fisher discriminant* and the distance of the labeled examples in the feature space. In [45] the RBF kernel is generalized with the *Anisotropic RBF* kernel. In particular, the RBF kernel defined as $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\beta_0 \|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$, can be considered as a special case of a more general $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j))$ where the matrix $\mathbf{M}$ has additional parameters to learn. In the Anisotropic RBF kernel we set $\mathbf{M} = \mathbf{M}_\beta = \mathbf{diag}(\beta^{(1)}, ..., \beta^{(m)})$ with $\beta^{(r)} \in \mathbb{R}_+$ the parameters.

This new formulation has a greater number of degrees of freedom than the classical RBF kernel. Finding the correct $\mathbf{M}_{\boldsymbol{\beta}}$ is a metric learning problem and corresponds to directly perform feature weighting on the original features.

2.2 Transductive feature extraction with non linear kernels

Algorithms in this second class of KL methods are able to perform feature extraction in the feature space defined by a non linear kernel. In this case, the feature mapping from the input space to the feature space is not explicitly defined. A popular solution is given by *Kernel Principal Component Analysis* (KPCA) [46]. KPCA uses the *kernel trick* to implicitly find the projections on the eigenvectors (principal directions) and the corresponding eigenvalues of the covariance matrix in feature space. KPCA can also be considered as an unsupervised manifold learning technique mapping data points into a lower-dimensional space.

The intrinsic problem of these algorithms is the transductive environment that they require (i.e. both training and test example needs to be available before training the classifier). This problem can be overcome by using *Out-Of-Sample* techniques to approximate the kernel values on new examples. Empirical experiments have shown that errors on examples in the *Out-Of-Sample* set and examples in the *In-Sample* set are similar [47].

2.3 Spectral kernel learning

These methods are founded on the spectral decomposition of the Laplacian graph $\boldsymbol{L}$, that is an undirected graph that contains the manifold structure of the data. By using these methods we are interested in finding the smoothest components of $\boldsymbol{L}$ (i.e. the eigenvectors with the smaller eigenvalues) and hence building a kernel which penalizes large changes between strongly connected nodes. This can be made by changing the spectral representation of $\boldsymbol{L}$ rescaling the eigenvalues in according to the semi-supervised information, for examples using linear programming [48]. More specifically, these algorithms are based on the possibility to write a semi-positive definite matrix $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ using the equation of its spectral decomposition: $\boldsymbol{L} = \sum_{s=1}^{n} \lambda_s \boldsymbol{u}_s \boldsymbol{u}_s^T$ where $\lambda_s \geqslant 0$ $\forall s = 1, \dots, n$ are the eigenvalues, with their corresponded eigenvectors $\{\boldsymbol{u}_s\}_{s=1}^{n}$. Then, the matrix $\boldsymbol{L}$ is a weighted sum of rank-1 matrices where the weights are the eigenvalues. A nonparametric spectral transformation $\tau : \lambda_s \rightarrow \mu_s$ optimized to the specific task is performed using the semi-supervised information contained in $\boldsymbol{L}$. Basically, a new set of features are created using the *spectral embedding*, and the matrix of the examples (where a row is an example) is now defined by $\mathbf{X} = \mathbf{U}\sqrt{\mathbf{D}_\mu} \in \mathbb{R}^{n\times n}$. In particular, the i^{th} example is defined by $\boldsymbol{x}_i = [\sqrt{\mu_s}\boldsymbol{u}_{i,s}]_{s=1}^{n}$ highlighting the relationship between feature learning and spectral kernel learning. Clearly, changing the eigenvalues, the algorithm is indirectly changing the weight of the features of the *spectral embedding.*

2.4 Multiple kernel learning (MKL)

MKL [49] is one of the most popular paradigm used to learn kernels in real world applications [50, 51]. The kernels generated by these techniques are combinations of previously defined *weak* kernels $\mathbf{K}_1, ..., \mathbf{K}_p$ with a constraint in the form: $H_p^q = \{x \mapsto \mathbf{w} \cdot \boldsymbol{\phi}_{\mathbf{K}}(x) : \mathbf{K} = \sum_{k=1}^{p} \mu_k \mathbf{K}_k, \boldsymbol{\mu} \in \Gamma_q, \|\mathbf{w}\| \leqslant 1\}$ with $\Gamma_q = \{\boldsymbol{\mu} : \boldsymbol{\mu} \succcurlyeq 0, \|\boldsymbol{\mu}\|_q = 1\}$ and considering the function $\boldsymbol{\phi}_{\mathbf{K}}$ as the feature mapping from the input space to the feature space. The value q being the kind of mean used, is typically fixed to 1 or 2.

These algorithms are supported by several theoretical results that bound the *estimation error* (i.e. the difference between the true error and the empirical margin error). These bounds exploit the *Rademacher complexity* applied to the combination of kernels [52, 53, 54]. Existing MKL approaches can be divided in two principal categories. In the first category, *Fixed or Heuristic*, some fixed rule is applied to obtain the combination. They usually get results scalable with respect to the number of kernels combined but their effectiveness will critically depend on the domain at hand. On the other hand, the *Optimization based* approaches learn the combination parameters by solving an optimization problem that can be integrated in the learning machine (e.g. structural risk based target function) or formulated as a different model (e.g. alignment, or other kernel similarity maximization) [55, 56, 57, 58].

The MKL optimization problem turned out to be a very challenging task as, for example, doing better than the simple average of the *weak* kernels is surprisingly difficult. Moreover, the *Optimization based* MKL algorithms have a high computational complexity. More recently, scalable methods have been proposed that can tackle thousands of kernels in a reasonable time and memory space [59]. For example, in [60], thousands of kernels can be combined using a fixed amount of memory and linear computation complexity with respect to the number of kernels. Having MKL algorithms which are scalable opens a new scenario for MKL. While standard MKL algorithms typically cope with a small number of strong kernels and try to combine them, a second perspective is also possible, that is, the MKL paradigm can be exploited to combine a very large amount of weak kernels, aiming at boosting their combined accuracy in a way similar to feature weighting.

3 Contributions to the ESANN 2015 Special Session on Feature and Kernel Learning

The *Feature and Kernel Learning* special session has received research works from four groups, presenting approaches to deal with feature selection and weighting, as well as the correct selection of kernels. Each accepted paper is briefly introduced in the following.

As discussed above, feature selection has proven effective in helping with the diagnosis of several diseases. In [61], the authors applied feature selection methods to improve the diagnosis of Evaporative Dry Eye (EDE), which is a prevalent disease that leads to irritation of the ocular surface, and that is associated with symptoms of discomfort and dryness. Existing approaches for the automatic classification of images to detect this disease have been focused on dark eyes, since they are most common in humans. The authors introduced also images from light eyes and presented a methodology making use of feature selection methods to learn which features are the most relevant for each type of eyes. The experimental results showed an improvement in the automatic classification of the tear film lipid layer, independent of the color of the eyes and with classification rates over 90%.

Traditionally, and because of the necessity of dealing with extremely high dimensional data, most of the novel feature selection contributions have fallen within the filter model. However, when the type of problem does not prevent its application, the wrapper model can obtain more powerful results. In [62], a feature selection wrapper is designed specifically for Echo State Networks. It

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

defines a feature scoring heuristics, that can be applied to a generic feature subset search algorithm, which allows to reduce the need for learning model retraining with respect to the wrappers found in the literature. The experimental assessment on real-word noisy sequential data shows that the proposed method can identify a compact set of relevant, highly predictive features with as little as 60% of the time required by the original wrapper.

In the real world problems there are latent variables that are not directly observable and the discovering of temporally delayed causalities is another issue for which a correct feature selection is fundamental. Specifically, only history-based features are able to represent these delayed causalities. In [63], a greedy algorithm (called *PULSE*) is presented to deal with this hard task in order to discover a sparse subset of features in a reinforcement learning scenario.

In the last few years, the specialized literature in feature selection has shown a tendency to mix algorithms, particularly in the form of ensemble learning. The rationale behind this approach is that, since no single method has demonstrated to be "the best", it might be better to rely on the output of several different methods. The authors in [64] propose a method to aggregate different ranking filters (into a better one) using Principal Component Analysis (PCA). The method presented is called *First Principal Component Projection Score* (FPCPS).

References

[1] Zheng Alan Zhao and Huan Liu. *Spectral feature selection for data mining.* Chapman & Hall/CRC, 2011.

[2] Isabelle Guyon. *Feature extraction: foundations and applications*, volume 207. Springer, 2006.

[3] Mark A Hall. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.

[4] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[5] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[6] Ian Jolliffe. *Principal component analysis.* Wiley Online Library, 2002.

[7] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[8] Jianping Hua, Waibhav D Tembe, and Edward R Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3):409–424, 2009.

[9] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13(1):27–66, 2012.

[10] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519, 2013.

[11] Harun Uğuz. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, 24(7):1024–1032, 2011.

[12] Hui-Huang Hsu, Cheng-Wei Hsieh, and Ming-Da Lu. Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, 38(7):8144–8150, 2011.

[13] Chien-Pang Lee and Yungho Leu. A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing*, 11(1):208–213, 2011.

[14] Juanying Xie and Chunxia Wang. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 38(5):5809–5815, 2011.

[15] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12):e28210, 2011.

[16] Feng Yang and KZ Mao. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(4):1080–1092, 2011.

[17] Jing Yang, Dengju Yao, Xiaojuan Zhan, and Xiaorong Zhan. Predicting disease risks using feature selection based on random forest and support vector machine. In *Bioinformatics Research and Applications*, pages 1–11. Springer, 2014.

[18] Sahand Khakabimamaghani, Farnaz Barzinpour, and Mohammad R Gholamian. Enhancing ensemble performance through feature selection and hybridization. *International Journal of Information Processing and Management*, 2(2), 2011.

[19] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Data classification using an ensemble of filters. *Neurocomputing*, 135:13–20, 2014.

[20] V Bolón-Canedo, N Sánchez-Maroño, A Alonso-Betanzos, JM Benítez, and F Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135, 2014.

[21] Seung Ho Lee, Jae Young Choi, Konstantinos N Plataniotis, and Yong Man Ro. Color component feature selection in feature-level fusion based color face recognition. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–6. IEEE, 2010.

[22] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian. Feature selection using principal feature analysis. In *Proceedings of the 15th international conference on Multimedia*, pages 301–304. ACM, 2007.

[23] Aouatif Amine, Ali El Akadi, Mohammed Rziza, and Driss Aboutajdine. Ga-svm and mutual information based frequency feature selection for face recognition. *GSCM-LRIT, Faculty of Sciences, Mohammed V University, BP*, 1014, 2009.

[24] Rabab M Ramadan and Rehab F Abdel-Kader. Face recognition using particle swarm optimization-based selected features. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2(2):51–65, 2009.

[25] Debasis Mazumdar, Soma Mitra, and Sushmita Mitra. Evolutionary-rough feature selection for face recognition. In *Transactions on rough sets XII*, pages 117–142. Springer, 2010.

[26] Hamidreza Rashidy Kanan and Karim Faez. An improved feature selection method based on ant colony optimization (aco) evaluated on face recognition system. *Applied Mathematics and Computation*, 205(2):716–725, 2008.

[27] E Ataer-Cansizoglu, J Kalpathy-Cramer, S You, K Keck, D Erdogmus, MF Chiang, et al. Analysis of underlying causes of inter-expert disagreement in retinopathy of prematurity diagnosis. *Methods Inf Med*, 54(1):93–102, 2015.

[28] B. Remeseiro, V. Bolon-Canedo, D. Peteiro-Barral, A. Alonso-Betanzos, B. Guijarro-Berdinas, A. Mosquera, M.G. Penedo, and N. Sanchez-Marono. A methodology for improving tear film lipid layer classification. *Biomedical and Health Informatics, IEEE Journal of*, 18(4):1485–1493, 2014.

[29] Michael C Lee, Lilla Boroczky, Kivilcim Sungur-Stasik, Aaron D Cann, Alain C Borczuk, Steven M Kawut, and Charles A Powell. Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. *Artificial intelligence in medicine*, 50(1):43–53, 2010.

[30] G Ilczuk, R Mlynarski, W Kargul, and A Wakulicz-Deja. New feature selection methods for qualification of the patients for cardiac pacemaker implantation. In *Computers in Cardiology, 2007*, pages 423–426. IEEE, 2007.

[31] Cheng-Lung Huang and Cheng-Yi Tsai. A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications*, 36(2):1529–1539, 2009.

[32] Kyoung-jae Kim and Ingoo Han. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications*, 19(2):125–132, 2000.

[33] Ming-Chi Lee. Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, 36(8):10896–10904, 2009.

[34] Chih-Fong Tsai and Yu-Chieh Hsiao. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, 50(1):258–269, 2010.

[35] George Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of ;achine Learning Research*, 3:1289–1305, 2003.

[36] Hyunsoo Kim, Peg Howland, and Haesun Park. Dimension reduction in text classification with support vector machines. In *Journal of Machine Learning Research*, pages 37–53, 2005.

[37] Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski, and Michael W Mahoney. Feature selection methods for text classification. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 230–239. ACM, 2007.

[38] George Forman. Feature selection for text classification. *Computational Methods of Feature Selection*, pages 257–276, 2008.

[39] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[40] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

[41] Gert R. G. Lanckriet, Nello Cristianini, Peter L. Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[42] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

[43] Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2760–2768. Curran Associates, Inc., 2013.

[44] Wenjian Wang, Zongben Xu, Weizhen Lu, and Xiaoyun Zhang. Determination of the spread parameter in the gaussian kernel for classification and regression. *Neurocomputing*, 55(3):643–663, 2003.

[45] Fabio Aiolli and Michele Donini. Learning anisotropic RBF kernels. In *Artificial Neural Networks and Machine Learning - ICANN 2014 - 24th International Conference on Artificial Neural Networks, Hamburg, Germany, September 15-19, 2014. Proceedings*, pages 515–522, 2014.

[46] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Artificial Neural Networks ICANN97*, pages 583–588. Springer, 1997.

[47] Yoshua Bengio, Jean-Francois Paiement, and Pascal Vincent. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *In Advances in Neural Information Processing Systems*, pages 177–184. MIT Press, 2003.

[48] Wei Liu, Buyue Qian, Jingyu Cui, and Jianzhuang Liu. Spectral kernel learning for semi-supervised classification. In *IJCAI*, pages 1150–1155, 2009.

[49] Mehmet Gönen and Ethem Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[50] Serhat S Bucak, Rong Jin, and Anil K Jain. Multiple kernel learning for visual object recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1354–1369, 2014.

[51] Alexander Zien and Cheng Soon Ong. Multiclass multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1191–1198. ACM, 2007.

[52] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 247–254, 2010.

[53] Zakria Hussain and John Shawe-Taylor. Improved loss bounds for multiple kernel learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 370–377, 2011.

[54] Marius Kloft and Gilles Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In *Advances in Neural Information Processing Systems*, pages 2438–2446, 2011.

[55] Alain Rakotomamonjy, Francis R. Bach, St'ephane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 208.

[56] Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM, 2009.

[57] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Non-sparse regularization and efficient training with multiple kernels. *CoRR*, abs/1003.0079, 2010.

[58] Zenglin Xu, Rong Jin, Haiqin Yang, Irwin King, and Michael R. Lyu. Simple and efficient multiple kernel learning by group lasso. In *ICML*, pages 1175–1182, 2010.

[59] A. Jain, S. V. N. Vishwanathan, and M. Varma. Spg-gmkl: Generalized multiple kernel learning with a million kernels. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, August 2012.

[60] Fabio Aiolli and Michele Donini. Easy multiple kernel learning. In *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014*, 2014.

[61] Beatriz Remeseiro, Verónica Bolón-Canedo, Amparo Alonso-Betanzos, and Manuel G. Penedo. Learning features on tear film lipid layer classification. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

[62] Davide Bacciu, Filippo Benedetti, and Alessio Micheli. Esnigma: efficient feature selection for echo state networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

[63] Robert Lieck and Marco Toussaint. Discovering temporally extended features for reinforcement learning in domains with dealyed causalities. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

[64] Andrey Filchenkov, Vladislav Dolganov, and Ivan Smetannikov. Pca-based algorithm for feature ranking filters ensembling. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.