Dimensionality Reduction: An Effective Technique for Feature Selection

Swati A Sonawale
PG Student
Department of computer Engineering
Dr.D.Y.Patil School of Engineering & Technology
Savitribai Phule Pune University
Pune,India

Roshani Ade
Assistant Professor
Department of computer Engineering
Dr.D.Y.Patil School of Engineering & Technology
Savitribai Phule Pune University
Pune,India

ABSTRACT
For knowledge gaining the dimensionality reduction is a significant technique. It has been observed that most of the time dataset is multidimensional and larger in size. When we are using same dataset for classification it may create wrong results and it may also requires more requirements in terms of storage as well as processing capability. Most of the features present are redundant, inconsistent and degrade the performance. To increase the effectiveness of classification these duplicate and inconsistent features must be removed. In this research we have introduced a new method for dealing with the problem of dimensionality reduction. By reducing the unrelated (irrelevant) and unnecessary features related to data, or by means of effectively merging original features to produce a smaller set of feature with more discriminative control, dimensionality reduction methods convey the instant effects of rapid the data mining algorithms, better performance, and increase in unambiguous of data model

Keywords
Dimension reduction, Fuzzy ARTMAP, Feature selection, Feature extraction, Supervised and Unsupervised techniques, semi-supervised techniques.

1. INTRODUCTION
The increase in information gaining capacity, cost decreasing in information storage and development of database and data store (warehouse) technology have led to the emergence of high dimensional dataset, from recent decades.

The data size increases in terms of number of features and number of instances becomes a provocation for the feature selection algorithms.

From these various features many features are redundant and irrelevant which increase the search space size which further causing in difficulty to process the data. This curse of dimensionality is a major obstacle within machine learning and application under data mining. To handle the high dimensional data, huge amount of storage planetary and computational time is required and thus dimensionality of data need to reduce. Whenever there is deal with high-dimensional data one significant way is Dimensionality reduction. Dimensionality reduction can be applied to decrease the dimensionality of the innovative data and increase in performance of machine learning techniques. The technique of dimensional reduction bring the immediate effects of speeding up data mining performance improvement and comprehensibility of models, by removing the unnecessary features or by successfully conjunction of original features to create set of features with more discriminative powers[1].

In paper [3], author proposed a novel algorithm for linear dimensionality reduction, called as Locality Preserving Projections (LPP). This LPP technique builds a graph including neighborhood data of the data set. The transformation matrix is computed by means of the concept of the Laplacian of the graph in which transformation matrix plots the data points to a subspace. In certain sense, this linear transformation preserves local neighborhood information.

Dimensionality reduction performed within two types of techniques firstly, Feature extraction and secondly, Feature selection. These two dimension reduction categories can be differentiated as: feature extraction techniques generates a small set of novel features by merging the original features, though feature selection techniques picks a small set of the original features.

Feature extraction (FE) methods can be categorized as supervised feature extraction and unsupervised feature extraction according to accessibility of label information. The standard FE algorithms are usually categorized into linear and nonlinear algorithms. The purpose of Linear algorithms are to project the data with high-dimensionality to a lower-dimensional planetary by linear transformations according to certain criteria, which can be listed: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Maximum Margin Criterion (MMC), etc. Besides, nonlinear algorithms aim to project the original information by nonlinear transformations though conserving particular local information rendering to Laplacian Eigen maps and ISOMAP. Some standards such as: Locally Linear Embedding (LLE).

Feature selection called as subclass selection is a method generally used in machine learning, wherein a small set of the features obtain from the data are choosed for use of a learning algorithm. Feature selection achieves dimensionality reduction by selecting a small set of the original features.

To achieve this, a feature evaluation criterion is used with a search strategy to identify the relevant features. Feature selection is type of dimensionality reduction which increases the interpretability of learning model by keeping the original features, which is desired in numerous real applications, like text mining which is part of data mining and genetic application analysis.

The feature selection and feature extraction selection are depend on the particular application domain and specific data set. Feature selection is a sound defined issue in machine learning and data mining communities, particularly within supervised and unsupervised paradigms, the topic of numerous mechanisms [5], [6].

The task of feature selection come to be very challenging with the small-labeled-sample issues, in which the amount of unlabeled information can be more larger than the amount of labeled data [13]. Supervised feature selection algorithms need a huge amount of training data which is labeled. As

effect, such algorithms offer insufficient data about the construction of the objective concept, and can therefore fail to recognize the applicable features that are discriminative to various classes. Whereas, unsupervised feature selection algorithms ignore label data and hence may lead to performance deterioration.

Using together labeled and unlabeled data is probable to improved estimate feature relevance, under the consideration that labeled and unlabeled data are sampled from the similar population generated by target concept. Utility of

Semi-supervised feature selection is much modified and its efficiency has been demonstrated.

Semi-supervised dimensionality reduction can be looks like a novel issue in semi-supervised learning, which studies from a mixture of both labeled and unlabeled data. In several real time applications of data mining, unlabeled training samples are willingly accessible but labeled ones are equally expensive to obtain, and hence semi-supervised learning has concerned much attention.

Recent research on semi-supervised learning could be roughly classified into three classes, i.e. semi-supervised classification [19], semi-supervised regression [17, 20], and semi-supervised clustering [16]. Advance techniques of semi-supervised learning can be found in an excellent recent survey [21].

Embedded model such as C4.5 and LARS [24], integrate feature selection as a category of the learning procedure, and utilize the objective property of the learning model to monitor searching for feature which are relevant. Weighting algorithms.

Feature selection has been applied to different zones as an important mechanism, including text mining [25], computer vision [23], and bioinformatics [30], etc.

Moreover, feature selection algorithms may return either one a small set of features [33] or the weights of all features algorithms mostly categorized into three models: filter, wrapper or embedded model [27].

From these the filter model calculates features deprived of involving some learning algorithm. The wrapper model needs a learning algorithm and performance utilization to calculate the best of features.

2. LITERATURE SURVEY

2.1 Feature Extraction

From previous study to available the label information, the feature extraction techniques can be categorized as supervised FE or unsupervised FE. One of the sample of supervised feature extraction methods is Fisher Linear Discriminant (FLD), which can extract the ideal discriminant vectors when class labels are available. Besides, unsupervised feature extraction methods, the popular Principal Component Analysis (PCA) attempts to reservation the comprehensive covariance structure of data when class labels are not available. Additional methods can be originate in the literature commerce with feature extraction, (LLE: Locally Linear Embedding), (k-PCA: Kernel PCA), (LE: Laplacian Eigen map) [4] and (LPP: Locality Preserving Projection) [3].

2.2 Feature selection

For many years, problems occurred in feature selection has been examined by the statistics and machine learning communities. Recently, it has received much concentration since study in data mining is enthusiastic. In machine learning and data mining organizations suffers addressed problems of feature selection, especially in supervised and unsupervised models, the topic of several mechanisms [5]. In the supervised context, the significance of a feature can be estimated by its relationship with the class label: Fisher score (FS), Relief and Relief [8], Fast Correlation-Based Filter (FCBF) [9], and Spectrum decomposition (SPEC) [10].

Due to the absence of class labels which would guide the relevant data search, the unsupervised feature selection is considered as a much more difficult problem. The unsupervised Variance score (VS), Laplacian score [11], SPEC [10], Hilbert-Schmidt Independence Criterion (HSIC) [12].

The choosing technique is an essential component of feature selection. Spectral feature selection learns how to choose features rendering the constructions of the graph encouraged from a set of pairwise instance similarity [10]. Spectral feature selection can take different forms which can be listed as: separability, data dependency, reliability, performance of learning model which used in the wrapper model, etc.

3. FEATURE EXTRACTION TECHNIQUE

3.1 PCA-based Feature Extraction

The objective of PCA is to search a subspace whose basis vectors parallel to the maximal variances directions [36]. The evaluation cost of PCA mostly lies within the Singular Value Decomposition (SVD). And yet PCA can search the more illustrative features, it disregards the valuable class label data and, hence, is not optimum for usual classification tasks. The PCA technique is an unsupervised technique meanwhile it does not consider the classes within the training dataset. Even though it is best for reconstruction, it is unnecessarily ideal from a discrimination point of view.

3.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is used to search a lower-dimensional space that finest discriminates the instances from various classes [36]. Its objective is to maximize the Fisher standard, i.e., an objective function: are known Interclass scatter matrix and Intraclass scatter matrix, correspondingly. LDA unambiguously utilizes the label data of the samples which is appropriate for categorization issues. Furthermore, it is quite problematic for LDA to handle databases with high-dimensional illustration or streaming information.

Though LDA methods are based on heavy expectations that might not hold in various presentations, it turned out that the linear discriminant functions can harvest satisfactory consequences even when the covariance constructions are dissimilar. Therefore, LDA methodologies have been effectively practical in numerous classification difficulties such as multimedia information retrieval, image recognition and medical applications.

3.3 Maximum margin criterion

Maximum Margin Criterion (MMC) [35], [36] is a currently proposed supervised FE algorithm. Established on the same illustration as LDA, MMC goals to maximize the objective function.

Even though both MMC and LDA are supervised subspace learning methods, the calculation of MMC is simpler than that of LDA meanwhile MMC doesn't have inverse matrix

procedure. Analogous to other batch feature extraction methodologies, MMC is inefficient for huge-scale data or streaming information issues. Still, this technique is sensitive to parameter situations, which is an exposed problem.

3.4 Orthogonal centroid algorithm

Orthogonal Centroid (OC) algorithm is a lately projected supervised FE algorithm which exploits orthogonal transformation on centroid. It has been confirmed to be very efficiently for categorization problems and is based on the computation on vector space in linear algebra by QR matrix decomposition.

The Orthogonal Centroid algorithm for dimensionality reduction has been effectively applied on text data [36]. Though, cost of the time and space of QR disintegration are moreover exclusive for large-scale data like Web documents.

4. FEATURE SELECTION TECHNIQUES

4.1 Forward feature selection

The forward feature selection process starts by estimating all feature subsets that consist of just single input attribute. Exhaustive search is employ to find the overall finest set of input feature.

Exhaustive search starts with finding the best one-component a small set of the features, which is the similar in algorithm called the forward selection algorithm; after that it goes to search the finest a small set of two-component feature which might contain the input features as pair.

4.2 RFS

One of the most important issues in classification is nothing but feature selection. Various filter and wrapper techniques have been projected. Random Feature selection (RFS) is an efficient technique which is based on R-value, which is a parameter that is used to detect the overlapped spaces between classes within a feature.

The difference between conventional Forward Selection (FS) and RFS is that at every step to add an extra feature into the small set, FS contemplates all the residual features, whereas RFS just attempts more promising part of them.

4.3 Greedy

The very core greedy algorithm for feature selection and for regression, matching pursuit, customs the correlation in between the remaining and the candidate features to choose which feature to add next. Below a convinced irrepresentable situation on the design matrix which is not dependent of the sparse target, the greedy algorithm can choose features reliably when the size of sample goes to infinity. Besides, below a sparse eigenvalue situation, the greedy algorithm can consistently identify features on condition that each nonzero coefficient is superior than a constant times the noise level.

4.4 Feature selection with Laplacian score

Laplacian Score technique is projected in paper [14] to select features that preserve locality of sample identified by an affinity matrix, its Laplacian matrix and corresponding degree matrix. Subsequently in Laplacian Score technique, features are independently evaluated, selection of particular features with Laplacian Score can be accomplished by greedily selecting the top particular features having the lowest score values. This score can be used for unsupervised feature selection. It prefers those features by means of larger differences which have more illustrative power. Moreover, it have a tendency to choose features with stronger ability for locality preservation. A core consideration in Laplacian score is that data from the same class are neighboring to each other.

4.5 Feature selection with SPEC

As proposed in [34], Spectral feature selection (SPEC) is an extension of Laplacian Score to create it additional robust to noise. In SPEC, the three evaluation standards are projected for calculation of feature relevance are the affinity matrix, the normalized Laplacian matrix and the degree matrix. The SPEC framework agrees for diverse similarity matrix parameters, and ranking function. SPEC can produce a range of spectral feature selection algorithms for together unsupervised also for supervised learning. Therefore, SPEC is an overall architecture for the purpose of feature selection [10].

4.6 Feature selection with fisher score

Fisher score is one of the feature selection technique which selects features that allocate like values to the samples feature from the similar class and unlike values to samples from dissimilar classes. In paper [28], it has been shown that Fisher Score is a distinctive instance of Laplacian Score.

4.7 Feature selection with trace ratio Criterion

Instead of measuring the scores of overall the subsets of feature, traditional approaches compute the score for every feature, and after that select the important features which based on the rank of level of feature scores. Though, choosing the feature subset created on the feature-level score can't assurance the finest of the subset-level score. Subset-level score is directly optimized which projected to a new algorithm to effectively search the global finest subset of features such that the subset-level score is increased. The trace ratio criterion for subset-level feature selection is also projected in [29].

4.8 Feature selection with relief

Relief and its multiclass extension Relief both are supervised feature weighting algorithms of the filter model. The assessment standards of Relief and Relief suggest that these two algorithms choose features contributing to the petitioning of samples from various classes.

4.9 Feature selection with HSIC

Hilbert-Schmidt Independence Criterion (HSIC) is initially proposed in [26] for calculating the interdependencies in between two kernels. In paper [32], HSIC is comprehensive and also applied to process of feature selection. The main hint is to choose a subset of features, such that the achieved kernel maximizes the HSIC criterion with respect to assumed kernel matrix.

5. ALGORITHM

Dimensionality reduction is a significant task when dealing with high dimensional data it can be applied to reduce dimensionality of the original data and improve learning performance. In feature selection it has been recognized that the combination of individually good features don't necessarily lead to good learning performance. In other words the m best features are not the best set of m ones. Instead redundant features unnecessarily Constraint selection algorithm play important role in the dimensionality reduction technique.

5.1 Constraint Algorithm

Step 1: Select Dataset from System.

D = { WAVE, …}

Step 2: Create and Initialize Instances.

X =x1, x2,..., xN}

Step 3: Divide instances into Categories i.e. labeled instance and unlabeled instances.

XL = {x1, x2,…, xN}

XU = {x1, x2,…, xN}

Step 4: Create and Initialize Constraints.

Ω= {(x1, x2), (x1, x3) … (L*(L-1)/2) }

Step 5: Divide constraints into Categories i.e. Must-Link Constraints and Cannot- Link Constraint.

Ω'ML = {Must-Link Constraints}

Ω'CL = {Cannot -Link Constraints}

Step 6: Calculate Global coherence of constraint Coh(Ω)

Step 7: for i=1 to Ω

if Coh(Wi) >= Coh(Ω) then

ΩS = ΩS U {Wi}

end if

end for

Selected Constraints are

ΩS = Ω'ML U Ω'CL

Following figure 1 shows flow of execution of constraint selection algorithm.

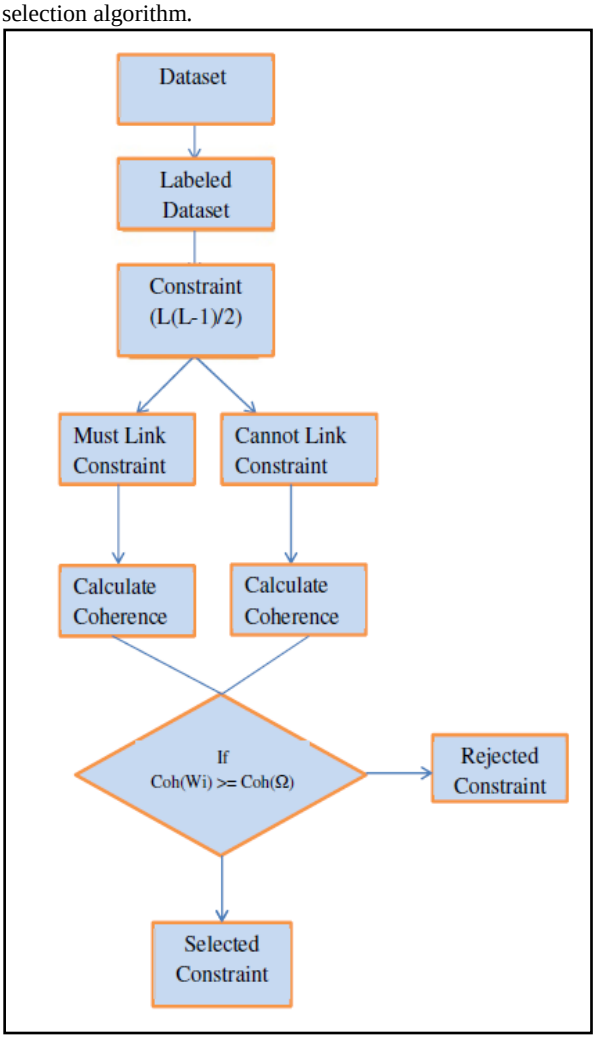

Fig. 1 Proposed System Flow

6. RESULTS

6.1 Constraint selection algorithm

By applying constraint selection algorithm we get following results as shown in table 1 & features which are selected are marked as a true and which are rejected are marked as a false.

i.e Coh(Wi) >= Coh(Ω) then

ΩS = ΩS U {Wi}

Table 1. Result of constraint selection algorithm

Sr.No	Coh(Wi)	Coh(Ω)	Status
1	0.777303234	0.640956891	True
2	0.73947529	0.640956891	True
3	0.730933496	0.640956891	True
4	0.749847468	0.640956891	True

5	0.640591966	0.640956891	False
6	0.765100671	0.640956891	True
7	0.573542736	0.640956891	False
8	0.584113561	0.640956891	False
9	0.653881003	0.640956891	True
10	0.776082977	0.640956891	True
11	0.761439902	0.640956891	True
12	0.761439902	0.640956891	True
13	0.55602537	0.640956891	False
14	0.572032619	0.640956891	False
15	0.785845027	0.640956891	True

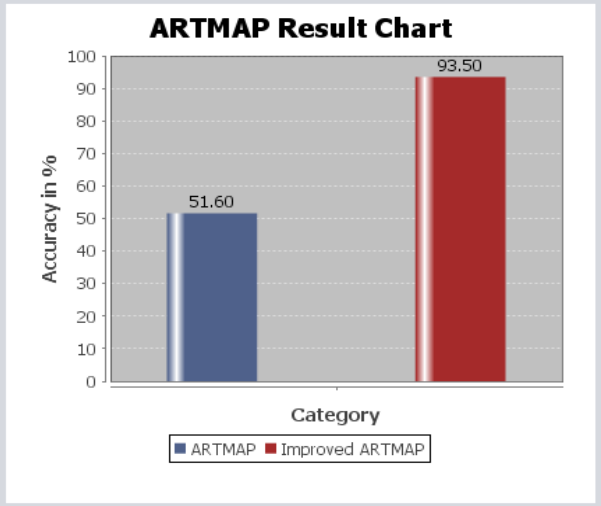

Figure 2.Result of constraint selection algorithm

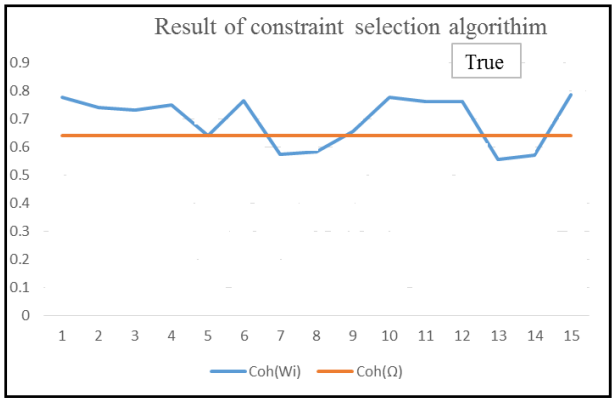

Figure 3. Comparision of ARTMAP and improved ARTMAP after reducing features

Total Constraint=4950
Selected constraint=3928

Above figure 3 shows comparision between ARTMAP & Improved ARTMAP algorithims in terms of classification accuracy. Due to dimensionality reduction accuracy is increased as well as it saves time.

7. CONCLUSION & FUTURE ENHANCEMENT

This framework for feature selection is based on constraint selection and redundancy elimination for semi-supervised dimensionality reduction. A new score function was developed to evaluate the relevance of features based on both, the locally geometrical structure of unlabeled data and the constraint preserving ability of labeled data.

In future we can use another classifier for classification so that it will save more time. Thus, Feature selection is an essential step in successful data mining applications which can effectively reduce data dimensionality by removing irrelevant features if we use that data our processing time will reduced.

8. ACKNOWLEDGMENT

Sincerely thank to all anonymous researchers for providing us such helpful opinion, findings, conclusions and recommendations. Also thank to guide Prof. Roshani Ade, HOD Prof .Arti Mohanpurkar, Principal Dr.Uttam Kalwane & colleagues for their support and guidance.

9. REFERENCES

[1] Z. Zhao and H. Liu, Spectral Feature Selection for Data Mining, USA: Chapman and Hall-CRC, 2012.

[2] I.. Jolliffe, Principal Component Analysis, USA: Springer, 2002.

[3] X. He and P. Niyogi, "Locality preserving projections," in Proc. NIPS, 2004.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in Proc. NIPS, 2002.

[5] I Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.

[6] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," J. Mach. Learn. Res., vol. 5, Aug. 2004, pp. 845–889.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, New York, NY, USA: Wiley Interscience, 2000.

[8] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of relief and relieff," Mach. Learn., vol. 53, no. 1–2, pp. 23–69, 2003.

[9] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," J. Mach. Learn. Res., vol. 5, Oct. 2004, pp. 1205–1224.

[10] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in Proc. 24th Int. Conf. Mach. Learn., Corvallis, OR, USA, 2007.

[11] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Proc. NIPS, Vancouver, Canada, 2005.

[12] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," J. Mach. Learn. Res., vol. 13, no. 1, Jan. 2012, pp. 1393–1434.

[13] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," in Proc. SIAM Int. Conf. Data Mining, Tempe, AZ, USA, 2007, pp. 641–646.

[14] D. Zhang, Z. Zhou, and S. Chen, "Semi-supervised Dimensionality reduction," in Proc. SIAM Int. Conf. Data Mining, Pittsburgh, PA, USA, 2007.

[15] O. Chapelle, B. SchÄolkopf, and A. Zien, editors. Semi-Supervised Learning. MIT Press, Cambridge, 2006.

[16] S. BASU, M. BILENKO, AND R. MOONEY, A probabilistic framework for semi-supervised clustering, in KDD'04, Seattle, WA, 2004, pp. 59–68.

[17] U. Brefeld, T. G¨A RTNER, T. SCHEFFER, AND S. WROBEL, Efficient co-regularized least squares regression, in ICML'06, Pittsburgh, PA, 2006, pp. 137–144.

[18] K. WAGSTAFF, C. CARDIE, S. ROGERS, AND S. SCHROEDL, Constrained k-means clustering with background knowledge, in ICML'01, Williamstown, MA, 2001, pp. 577–584.

[19] T. ZHANG AND R. K. ANDO, Analysis of spectral kernel design based semi-supervised learning, in NIPS 18, MIT Press, Cambridge, MA, 2006, pp. 1601–1608.

[20] Z.-H. ZHOU AND M. LI, Semi-supervised learning with co-training, in IJCAI'05, Edinburgh, Scotland, 2005.

[21] X. ZHU, Semi-supervised learning literature survey, Tech. Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006. http://www.cs.wisc.edu/»jerryzhu/pub/ssl survey.pdf.

[22] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. John Wiley & Sons, New York, 2 edition, 2001.

[23] J. G. Dy and etal. Unsupervised feature selection applied to content-based retrieval of lung images. Transactions on pattern Analysis and Machine Intelligence, 25(3):373–378, 2003.

[24] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. Annals of Statistics, 32:407–49, 2004.

[25] G. Forman. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3:1289–1305, 2003.

[26] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with hilbert-schmidt norms. In Proceedings of ALT, 2005.

[27] Guyon and A. Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.

[28] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In Advances in Neural Information Processing Systems 18, 2005.

[29] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan. Trace ratio criterion for feature selection. In Proceedings of Conference on Artificial Intelligence (AAAI), 2008.

[30] Y. Saeys and etal. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19):2507–2517, 2007.

[31] M.R. Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and Relief. Machine Learning, 53:23–69, 2003.

[32] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. Journal of Machine Learning Research,

[33] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the ICML, 2003.

[34] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the ICML, 2007.

[35] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," Proc. Conf. Advances in Neural Information Processing Systems, 2004.

[36] Jun Yan, Benyu Zhang, Ning Liu, Shuicheng Yan, Qiansheng Cheng, Weiguo Fan, Qiang Yang, Wen si Xi, and Zheng Chen, "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing", Mar. 2006