2014 IEEE Congress on Evolutionary Computation (CEC)
July 6-11, 2014, Beijing, China

Feature Extraction Based on Trimmed Complex Network Representation for Metabolomic Data Classification

Yue Chen, Zexuan Zhu and Zhen Ji
Shenzhen City Key Laboratory of Embedded System Design,
College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen, China 518060

***Abstract*—Over the last few decades, metabolomics has been widely used to reveal the linkages between metabolite signal levels and physiological states. Metabolomic data are naturally high dimensional and noisy, which poses computational challenges for data analysis. In this study, a novel feature extraction method based on trimmed complex network representation is proposed for metabolomic data classification. Particularly, the proposed method begins with feature selection on the original data, and then a complex network of the selected features is constructed to represent each data sample. Afterward, the network edges are trimmed and a few topological network metrics are extracted as new features for the classification of the samples. The experimental results on a real-world metabolomic data of clinical liver transplantation demonstrate the efficiency of the proposed feature extraction method.**

I. Introduction

Metabolomics is an emerging field attracting great attention in the last few years. It is a scientific study of biochemical processes involves metabolites for understanding the fundamental of many diseases and the related metabolic responses [1]. Metabolite profiling is one of the most important research areas in metabolomics dedicated to reveal the linkages between metabolite signal levels and physiological states [2]. The metabolite profiling data (referred as metabolomic data in this study), mainly generated with mass spectrometry, chromatographic, or nuclear magnetic resonance spectroscopy technology, captures thousands of metabolite signal levels of a tissue in a specific physiological state, but only a small number of them show relevance to the specific physiological state and the others are noise [3], [4]. The instrument-dependent high dimensional and noisy nature of the data poses challenges for the computational analysis tasks like regression, classification, and clustering [5], [6]. Feature selection, weighting and/or extraction methods have been applied to filter the noises and improve the analysis accuracy [7]–[9].

This work was supported in part by the National Natural Science Foundation of China Joint Fund with Guangdong, under Key Project U1201256, the National Natural Science Foundation of China, under Grants 61171125 and 61205092, the Guangdong Foundation of Outstanding Young Teachers in Higher Education Institutions, under grant Yq2013141, Guangdong Natural Science Foundation under grant S2012010009545, the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China, under grand 20111568, and Shenzhen Scientific Research and Development Funding Program under grants ZYC201105170243A, KQC201108300045A and JCYJ20130329115450637.

All correspondence should be addressed to Prof. Zhen Ji (Email: jizhen@szu.edu.cn, Tel: +86 755 2655 7413).

Recently, a novel feature extraction method based on complex network representation was proposed for the classification of metabolomic data [10]. Particularly, the method transforms each data sample into a network whose vertexes represent the metabolomic spectral bins, i.e., the features, and the edges represent their intensities of associated with a disease. From the network constructed, a few topological network metrics are extracted to represent the samples. A feature selection based preprocessing method was further introduced to the extraction method by the same author [11], in order to reduce the network size and improve the classification accuracy. The feature extraction method not only leads to accurate classification thanks to its robustness to data noise, but also provides a unique informative perspective to understand the data samples. Yet, the feature selection used in [11] is unsupervised, i.e., the class label is not considered, which might miss out some important features. Moreover, the complex networks constructed in [10], [11] still contain noisy connections which would affect the metric extraction and the analysis of the data.

To overcome the problems, we introduce supervised feature selection and edge trimming methods to the complex network representation based feature extraction. Particularly, the importance of the features, i.e., the network vertexes, and the edges are evaluated based on mutual information (MI) and conditional information (CI), respectively, considering the target class label. Noisy or irrelevant vertexes and edges are trimmed out of the network according to the importance of the vertexes and edges, so that more accurate metric features can be extracted from the network. The new feature extraction method combining the new supervised feature selection and edge trimming methods is tested on a real-world metabolomic data of clinical liver transplantation. The experimental results demonstrate the efficiency of the proposed method.

The remainder of this paper is organized as follows. Section II describes the proposed feature extraction method based on trimmed complex network representations. Section III presents the experimental results of the proposed method on the real-world metabolomic data set. Finally, the conclusion is given in Section IV.

II. Methodology

In this section, we introduce the proposed feature extraction method based on trimmed complex network constructions for metabolomic data classification. As shown in Fig. 1, the

proposed method consists of four steps, i.e., feature selection, network construction, edge trimming and network metrics extraction. In the first step, feature selection selects important and relevant features based on MI measure, considering the target class label. In the second step, a complex network of the selected features is constructed to represent each sample. In the third step, irrelevant edges are trimmed out from each network based on CI measure also considering the target class label. Finally, in the last step, topological network metrics are extracted as new features of the samples. The topological network metrics are inputted to standard classifiers for the classification of the data. The details of the steps are provided in the following subsections.

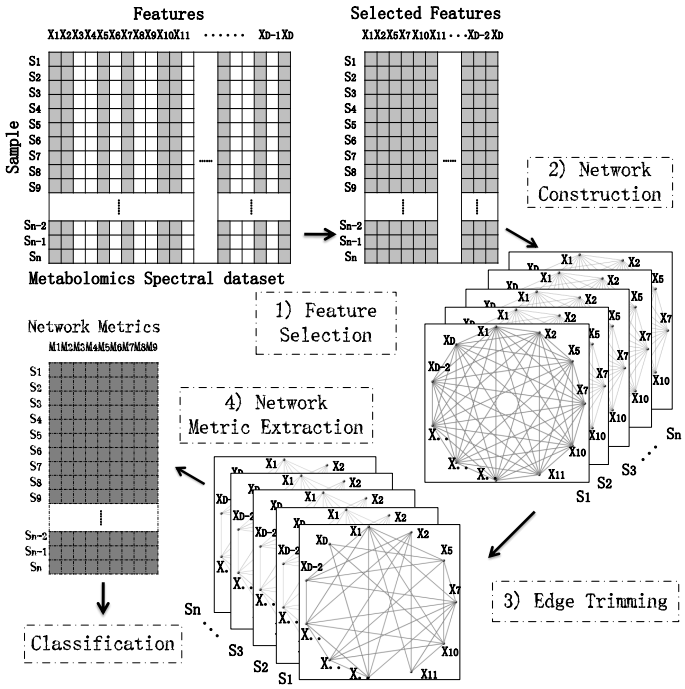


Fig. 1: The procedure of the feature extraction method

A. Supervised Feature Selection Based on Mutual Information (SMI)

Feature selection aims to select a small subset of K features from the original $D(D >> K)$ features so that the learning performance of the data is improved or not substantially deteriorated [12]. In [11], the relevance of each pair features is evaluated based on MI measure and the most correlated or uncorrelated features are selected. In this way, the feature selection is merely based on the relevancy between features, but the learning target is not involved. As a result, the selected features might not suit the learning task. Instead, in this study, supervised feature selection considering the class label based on MI is used. The relevance of the features to the class label is evaluated based on MI as follows:

$$I(X;C) = \sum_{x\in X}\sum_{c\in C} p(xc)\log\frac{p(xc)}{p(x)p(c)} \tag{1}$$

where X is a feature and C is the target class label vector; x and c are random variables in X and C, respectively; $p(x)$ and $p(c)$ are the two probability distribution functions; $p(xc)$ is the joint probability distribution function of x and c. For each feature X, an MI value $I(X;C)$ is calculated as a metric to assess the relevance of this feature to the class labels. In the principle of max-relevance [13], we sort the features in descending order according to $I(X;C)$ and select the top K features for network construction.

B. Network Construction

After feature selection, a undirected complex network of the K selected features is constructed to represent each sample. In mathematics, a network (or graph) is a representation of a set of objects (or vertexes) where some pairs of objects are connected by links (or edges) [14]. A complex network is a graph with non-trivial topological features. Complex networks have been used to characterize and analyze complex systems including metabolomic samples [15]. In the network representations of a metabolomic spectral data, each sample is represented by a network with each vertex being one of the available features, and the edges between two vertexes identifying exhibit characteristics related to the two features.

This study focuses on two-class classification problems of metabolomic data, where the class labels are either positive or negative. Following [10] and [11], to represent a sample with a undirected network, each selected feature is presented as a vertex, and every two vertexes are connected with a weighted edge associated with the normalized probability of the sample being positive when considering only the two corresponding features.

Let X and Y be two selected features, the normalized probability of a sample being positive in XY-space is estimated in the following three steps:

1) Linear regression. As shown in Fig.2, the green squares represent the positive samples (X_p, Y_p) and red circles represent the negative samples (X_n, Y_n) in XY-space. Two dashed lines are lineally fit to each group according to the following definitions:
$$\widetilde{Y_p} = \alpha_p^1 X_p + \alpha_p^0 \quad \widetilde{Y_n} = \alpha_n^1 X_n + \alpha_n^0 \tag{2}$$
where α_p^1 , α_p^0, α_n^1 and α_n^0 are coefficients of lines obtained with linear regression. $\widetilde{Y_p}$ and $\widetilde{Y_n}$ are approximate values of Y given X_p and X_n, respectively.
2) Estimating possibilities of a sample being positive and negative. Given a data sample S located in (x, y) in the XY-space, like the blue triangle plotted in Fig.2, the distances of the sample to the two dash lines are calculated as the two arrows in red and green shown. The probabilities of the sample S being positive $P_p(S)$ and negative $P_n(S)$ are estimated as follows:
$$\begin{aligned} P_p(S) &= Guass(\widetilde{y_p} - y, E(\Delta Y_p), D(\Delta Y_p)) \\ P_n(S) &= Guass(\widetilde{y_n} - y, E(\Delta Y_n), D(\Delta Y_n)) \end{aligned} \tag{3}$$
where $\widetilde{y_p}$ and $\widetilde{y_n}$ are approximate values y of S being a positive and negative sample according to (2), $\Delta Y_p =$

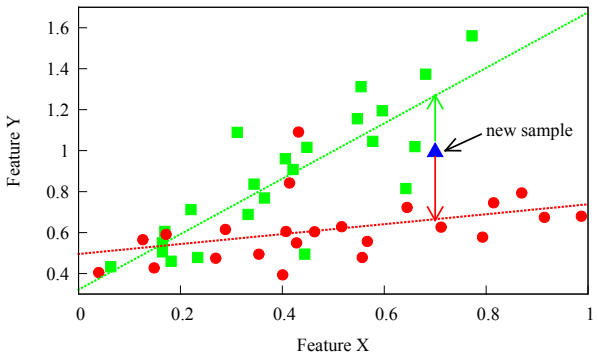


Fig. 2: Estimating edge weight of two features

$\widetilde{Y_p} - Y_p$, $\Delta Y_n = \widetilde{Y_n} - Y_n$, and $Guass(V, E(\cdot), D(\cdot))$ is a gaussian curve membership function evaluated at V with mean $E(\cdot)$ and variance $D(\cdot)$ of V.

3) Possibility normalization. The weight of the edge between feature X and Y is the normalized probability of the sample being positive, which is calculated as follows:

$$P(S) = \frac{P_p(S)}{P_n(S) + P_p(S)}$$

where $P_p(S)$ and $P_p(S)$ are the probabilities of the sample S being positive and negative calculated in (3).

After getting the normalized probabilities of pairwise features, we can construct the fully connected network (complete graph) with K vertexes each being a selected feature and $E = K * (K-1)/2$ edges each assigned a weight of the corresponding normalized probability.

C. Edge Trimming Based on Conditional Information (ETCI)

The constructed network contains many noise or trashy edges, which should be trimmed off to avoid disturbance of feature extraction. We proposed specific trimming method based on CI. A threshold P is introduced to decide how many percent of edges should be remained in the networks.

In this method, the edges are selected based on their pairwise relevance given the class labels. Particularly, the importance of the edges to the class label is evaluated based on CI as follows:

$$I(X;Y|C) = \sum_{c\in C} p(c) \sum_{x\in X} \sum_{y\in Y} p(xy|c) \log \frac{p(xy|c)}{p(x|c)p(y|c)} \tag{4}$$

where X and Y are two features connected by an edge, and C is the target class label vector. $p(xy|c)$ is the conditional joint probability distribution function. $p(x|c)$ and $p(y|c)$ are conditional probabilities distribution functions, respectively. $I(X;Y|C)$ denotes the information of the two features X and Y knowing the distribution of C. For each edge, we get a CI value $I(X;Y|C)$ as a metric to assess the relevance of this edge to the class labels. The edges are sorted in descending order based on $I(X;Y|C)$, and the top P percent of edges are maintained, i.e., the other edges are removed from the network.

After edge trimming, there are K vertexes and $P * E$ edges in each constructed network. And each edge is associated a weight indicating the normalized probability of sample being positive. From here, edges with small weights are further trimmed to simplify the network. Following [11], a edge weight threshold denoted as T is introduced and the edges with weights smaller than T are removed from the network. Setting P and T in $[40\%, 60\%]$ and $[0.4, 0.6]$, respectively, is observed in our empirical study to obtain satisfactory performance.

D. Network Metrics Extraction

After edge trimming, a few topological network metrics are extracted from a network as the new features of the corresponding sample. The network metrics listed below are considered in this study. The definitions of these metrics can be found in [14], [15].

- Link Sum
- Maximum degree
- Entropy of the degree distribution
- Link density
- Clustering coefficient
- Efficiency
- The number of components
- Max components size
- Entropy of eigenvector centrality distribution

E. Classification

The classification of the data samples is performed on the extracted new features using support vector machine (SVM) [16] implemented in LIBSVM [17]. Ten-fold cross-validation is used to evaluate the classification accuracy.

III. Experiments and Results

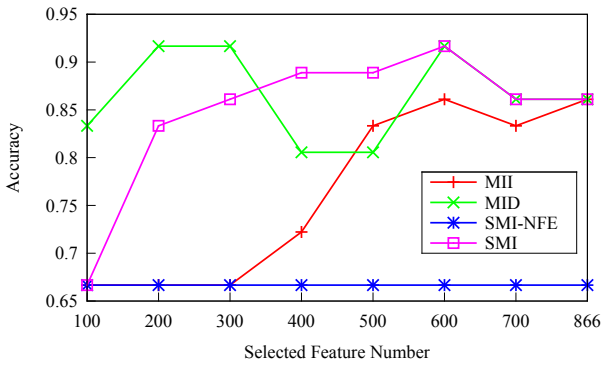
To test the performance of the proposed method, a real-world microdialysis-HPLC derived metabolomic data of 41 liver transplantation samples [18] is used. The data measures and records metabolomics mean levels at the donor and back table stages, and between early (2-6 h) and late (43-48 h) post-reperfusion. Each sample is detected in the chromatographic and the metabolite signal is split along the retention time into 866 features. In the following experiments, three clinical states are considered as class labels:

- Type of donor (TOD): livers used in transplantation are obtained from two types of donor, i.e. brain death donor (DBD) and cardiac death donor (DCD).
- Overall cold ischaemia time (OCIT): the sum of cold ischaemia time before the end of cold phase and cold ischaemia time elapsed before cold phase biopsy. According to OCIT, patients samples could be categorized into two groups, i.e., $OCIT \leq 471\ minutes$ and $OCIT > 471\ minites$.
- Peak Aspertate Aminotransferase (P-AST): P-AST is used to classify the extent of preservation injury after transplantation. According to the peak levels of serum AST during the first 72h after transplantation,

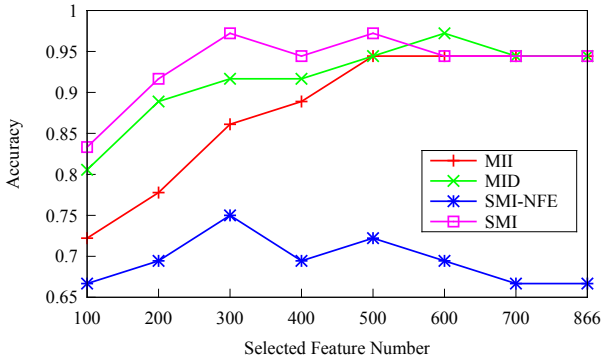
patients samples can be categorized into two groups, i.e., $P\text{-}AST \leq 3000\ U/L$ and $P\text{-}AST > 3000\ U/L$.

To study the effect of feature selection, in the first comparison study, the proposed supervised MI feature selection together with complex network representation but not using edge trimming, or SMI for short, is pitted against the feature selection methods MID and MII proposed in [11]. MID and MII are similar to SMI except that they select features with unsupervised MI in decreasing and increasing order, respectively. We also compared the performance of SMI to pure MI feature selection on the original data, i.e., no feature extraction SMI-NFE, to see the effect of network metric extraction. Experimental results of all algorithms for selecting different number of features are shown in Fig. 3. It is shown that methods based on feature extraction all have higher accuracy than SMI-NFE, which reveals that network representation is an efficient way for extracting more discriminative features. SMI show better performance than MID and MII, especially on studying OCIT and P-AST. From Fig. 3, it is also shown that 300 selected features i.e., $K = 300$ is suitable for studying the data.

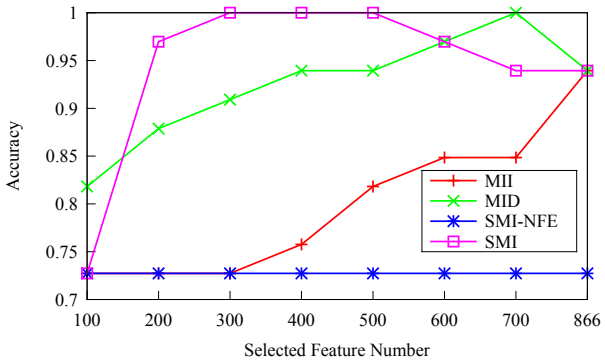
Edge trimming could reduce the impact of noise to the classification results as well as reduce the amount of computational time. To study the effect of edge trimming, we conduct edge trimming using all features (i.e., no feature selection) with different network weight threshold T, denoted as ETCI. And to study the effect of both feature selection and edge trimming, we use the algorithm with both feature selection method SMI and edge trimming method ETCI, denoted as SMI-ETCI. The results of SMI and N-FS-ET (i.e., not using neither feature selection or edge trimming) are also included for comparison. The results reported in Fig.4 show that edge trimming is capable of improving the final classification accuracy without doing feature selection, i.e., ETCI is better than N-FS-ET. The combination of SMI and ETCI, i.e., SMI-ETCI, taking advantage of the two components, performs better than the single SMI or ETCI.



(a) TOD



(b) OCIT



(c) P-AST

Fig. 3: Feature selection results. MII, MID, and SMI are feature selection based on network metric extraction with threshold $T = 0.5$. SMI-NFE is a pure feature selection method without using network metric extraction.

IV. Conclusion and Future Works

This paper proposed a feature extraction for metabolomics data classification based on trimmed complex network representations. We proposed supervised feature selection based on mutual information and edge trimming based on conditional information to simply the complex network representations. The experimental results on real-world data show the efficiency of the proposed feature extraction method. Nevertheless, we note that the propose method still have a few drawbacks to be addressed in the future work. For example, there should be a self-adaptive way to decide the parameters including the number of selected features K, the edge weight threshold T, and P the percent of edges should be remained.

References

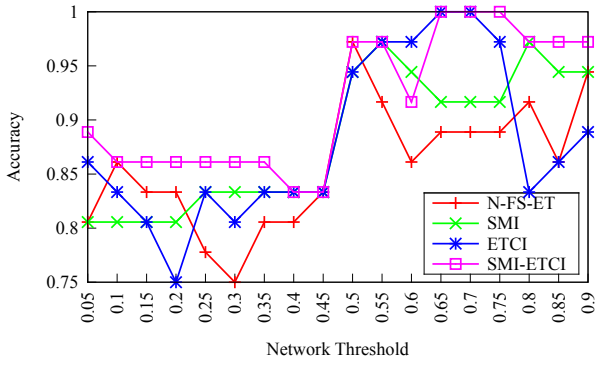
[1] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: acquiring and understanding global metabolite data," *TRENDS in Biotechnology*, vol. 22, no. 5, pp. 245–252, 2004.

[2] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer, "Metabolite profiling for plant functional genomics," *Nature biotechnology*, vol. 18, no. 11, pp. 1157–1161, 2000.

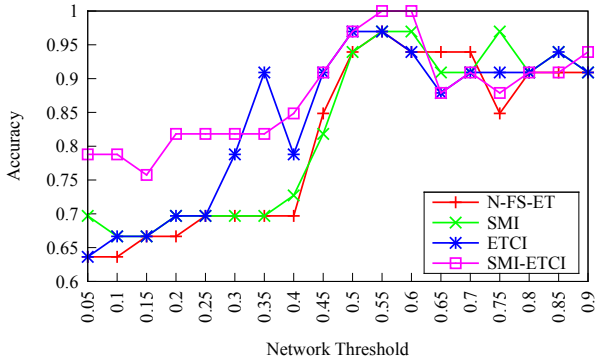
[3] M. Brown, W. B. Dunn, P. Dobson, Y. Patel, C. Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst, A. Tseng *et al.*, "Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics," *Analyst*, vol. 134, no. 7, pp. 1322–1332, 2009.

[4] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "Xcms: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787, 2006.

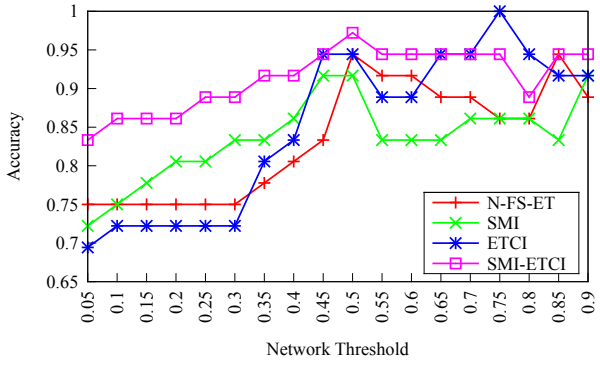
[5] H. Redestig and I. G. Costa, "Detection and interpretation of metabolite–



(a) TOD



(b) OCIT



(c) P-AST

Fig. 4: Performance of feature selection and/or edge trimming. All compared methods are based on network metric extraction, but N-FS-ET does not use feature selection or edge trimming. SMI is feature selection with $K = 300$ but without edge trimming. ETCI uses edge trimming $P = 50\%$ but no feature selection. SMI-ETCI uses both feature selection with $K = 300$ and edge trimming with $P = 50\%$.

transcript coresponses using combined profiling data," *Bioinformatics*, vol. 27, no. 13, pp. i357–i365, 2011.

[6] D. Brougham, G. Ivanova, M. Gottschalk, D. Collins, A. Eustace, R. O'Connor, and J. Havel, "Artificial neural networks for classification in metabolomic studies of whole cells using 1 h nuclear magnetic resonance," *Journal of Biomedicine and Biotechnology*, vol. 2011, 2010.

[7] R. A. Davis, A. J. Charlton, S. Oehlschlager, and J. C. Wilson, "Novel feature selection method for genetic programming using metabolomic ^{1}h nmr data," *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 1, pp. 50–59, 2006.

[8] K. Bryan, L. Brennan, and P. Cunningham, "Metafind: A feature analysis tool for metabolomics data," *BMC Bioinformatics*, vol. 9, no. 1, p. 470, 2008.

[9] S. Mahadevan, S. L. Shah, T. J. Marrie, and C. M. Slupsky, "Analysis of metabolomic data using support vector machines," *Analytical Chemistry*, vol. 80, no. 19, pp. 7562–7570, 2008.

[10] M. Zanin, D. Papo, J. L. G. Solís, J. C. M. Espinosa, C. Frausto-Reyes, P. P. Anda, R. Sevilla-Escoboza, R. Jaimes-Reategui, S. Boccaletti, E. Menasalvas *et al.*, "Knowledge discovery in spectral data by means of complex networks," *Metabolites*, vol. 3, no. 1, pp. 155–167, 2013.

[11] M. Zanin, E. Menasalvas, S. Boccaletti, and P. Sousa, "Feature selection in the reconstruction of complex network representations of spectral data," *PloS ONE*, vol. 8, no. 8, p. e72045, 2013.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[13] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[14] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4, pp. 175–308, 2006.

[15] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, p. 47, 2002.

[16] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[17] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[18] D. Richards, M. Silva, N. Murphy, S. Wigmore, and D. Mirza, "Extracellular amino acid levels in the human liver during transplantation: a microdialysis study from donor to recipient," *Amino Acids*, vol. 33, no. 3, pp. 429–437, 2007.