Using the Machine Learning Approach to Predict Patient Survival from High-Dimensional Survival Data

Wenbin Zhang
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada
wenbin.zhang@mun.ca

Jian Tang
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada
jian@mun.ca

Nuo Wang
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada
nw5516@mun.ca

***Abstract*—Survival analysis with high-dimensional data deals with the prediction of patient survival based on their gene expression data and clinical data. A crucial task for the accuracy of survival analysis in this context is to select the features highly correlated with the patient's survival time. Since the information about class labels is hidden, existing feature selection methods in machine learning are not applicable. In contrast to classical statistical methods which address this issue with the Cox score, we propose to tackle this problem by discretizing the survival time of patients into a suitable number of subgroups via silhouettes clustering validity. To cope with patients' censoring, we use "k-nearest neighbor" based on clinical parameters. Feature selection is then accomplished using Fast Correlation-Based Filtering approach from machine learning community. The effectiveness and efficiency of the proposed method are demonstrated through comparisons with classical statistical methods on real-world datasets and simulation datasets.**

Keywords—Survival prediction; machine learning; statistical method; high-dimensional survival data

I. INTRODUCTION

While a certain form of cancer is often thought of as a single disease, growing evidence suggests that there are multiple subtypes of a specific cancer that occur with clinically significant differences in survival [1]. One possible explanation is that two seemingly alike tumors are actually completely different diseases at the molecular profile of the tumor [2, 3, 4]. With the aim to ultimately improve the clinical management of cancer disease, researchers have sought to specify the subtypes of newly diagnosed patients, especially when those subtypes are associated with patient survival time, or elicit different prognoses and responses to certain therapies.

After collecting the survival information of a group of patients with the same cancer diagnosis, the survival prognosis can be predicted by studying the patient's survival profiles. Figure 1 illustrates the survival information obtained from the application of the method proposed in this work to the lung cancer dataset of Beer et al. [5]. As can be seen in Figure 1.a, patients with this type of cancer are at a high risk with the median survival time around 30 months. This type of fatal cancer must be treated aggressively, although aggressive treatments have potentially serious side effects. However, Figure 1.b indicates that there exists another subtype of this cancer, which is distinguished by a difference at the molecular level of the tumor and has a considerably improved long-term survival rates, with a median survival time of around 55 months. Patients with this less aggressive cancer can be treated with milder medications and still have excellent outcomes.

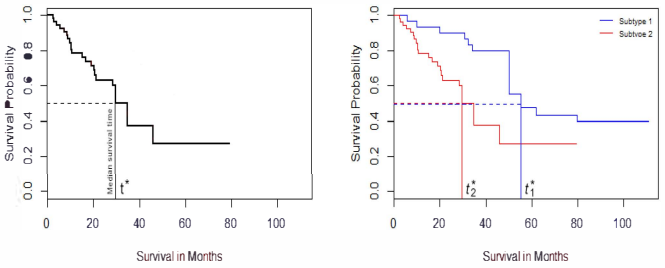

FIGURE 1 KAPLEN-MEIER SURVIVAL PLOTS FOR (A) ONE IDENTIFIED SUBGROUP AND (B) BOTH STRATIFIED SUBGROUPS OF THE LUNG CANCER DATASET

Cancer subtypes provide clues into patient disparities with respect to survival time, and can help in designing more targeted treatment strategies and more effective therapies. In recent years, a number of methods have been proposed for identifying cancer subtypes. These methods fall into three broad categories. Unsupervised learning techniques use only microarray data to unseal the concealed gene expression profile structure that defines cancer subtypes [8, 9]. Supervised learning approaches partition patients into subgroups based on clinical data exclusively [10, 11]. The exclusive use of either microarray data or clinical data in unsupervised and supervised methods for the identification of subgroup inhibits both the discovery of biologically meaningful subgroups and the likelihood of accurately predicting survival outcome. To overcome these difficulties, semi-supervised approaches [5, 7, 14, 15] are proposed that use both microarray and clinical data to guide the determination of cancer subtypes via the identification of biologically meaningful subgroups that are

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

correlated with clinical outcome. Bair and Tibshirani [7] proposed a semi-supervised approach that first used univariate Cox scores [12] to select genes that were highly correlated with the patients' survival time. Based on the selected genes, it applied unsupervised clustering techniques to partition the patients into groups, and used "nearest shrunken centroids" method [13] to derive prediction functions. Beer et, al [5] used similar idea to select genes, but proposed notion of 'risk index' as prediction function.

To the best of our knowledge, almost all the methods proposed so far on identifying subtypes associated with survival fall into the class of statistical procedures [5, 6, 7]. There is rarely a study using the techniques proposed by computer scientists in the machine learning community. Statistical procedures have achieved varying degrees of success. One of the shortcomings of these methods, however, is that they do not handle redundancy effectively. It is well known that in high-dimensional space like microarray data, redundant features can jeopardize generalization capability [4]. Moreover, selecting relevant features by iteratively fitting a univariate Cox proportional hazard model is time-consuming, especially in high-dimensional setting. These spark our interest in exploring the feasibility of applying the machine learning approach, in the context of survival data, to identify subtypes of cancer and to use this knowledge to diagnose future patients. Compared with the methods from statistical perspective, feature selection in machine learning possesses not only high relevancy, but also low redundancy with respect to the phenotype of interest. Furthermore, most of them are model free, which allows for wide applicability and easy implementation. This paper describes a new approach that utilizes both gene expression data and clinical data to conduct feature selection and survival prediction from the machine learning perspective. In our method, we first partition patients into separate groups using a sophisticated discretization method on their survival time. Then we use a highly efficient and effective feature selection method from machine learning community, Fast Correlation-Based Filter, to collect features that are correlated with survival time. Finally, we use typical machine learning algorithms to predict the patients' outcome. The experimental results confirm the effectiveness and efficiency of our method.

The rest of this paper is organized as follows. In Section 2, we describe the detail of our method. In Section 3, we use the experimental results to show the effectiveness and efficiency of our method. In Section 4, we conclude the paper by summarizing the main results.

II. Using Machine Learning Approach for High-Dimensional Survival Data

A. Coping with censoring

Classical statistical methods account for censoring with the Cox model, which keeps censored individuals in the risk set along with other individuals who have not yet experienced the event of interest [12]. In preference to standard statistical methods, we use k-nearest neighbor based on clinical parameters that are correlated to survival time to cope with patients' censoring. These clinical parameters are selected using penalized logistic regression and the penalized proportional hazard model with the Expectation Maximization (EM) algorithm [16]. They are then used to estimate censored survival time. The penalized complete log likelihood function can be written as

$$l(\gamma,\beta;\odot) = \sum_{i=1}^{n}(1-y_i)log[\,p(z_i)] + y_i log[1-p(z_i)] + \sum_{i=1}^{n} y_i\delta_i log[\,h(t_i|Y=1,x_i)] + y_i log[S(t_i|Y=1,x_i)] + n\lambda_{1j}\textstyle\sum_{j=1}^{q}|\gamma_j| + n\lambda_{2k}\sum_{k=1}^{m}|\beta_k| \quad (1)$$

where $p(t|z)$ is the probability of being cured given a covariate vector $z = (z_1,\ldots, z_q)'$, $S(t|x)$ is the survival function for uncured patients, conditional on $x = (x_1,\ldots, x_m)'$, and $\odot = \mathrm{p}(\mathrm{t_i}, \delta_i, \mathrm{x_i}, \mathrm{z_i}, \mathrm{y_i})$ represents the complete data for the ith individual with δ_i denoting censoring indicator and y_i the indicator of cure status, $i = 1,\ldots,n$; $\lambda|\cdot|$ is the lasso penalty function [17] and $\lambda = (\lambda_{11},\ldots\lambda_{1q},\lambda_{21},\ldots\lambda_{2m})$ is the tuning parameter, which can be chosen via GCV [18]. If prior biological knowledge shows that a certain variable has a known involvement in the cancer process, we can remit the penalty on the variable by setting the corresponding tuning parameter in λ to zero.

After a list of significant clinical parameters has been compiled, we compute the survival time for censored patients with the selected clinical covariates with regards to the proximities among them. The definition of "proximity" we employ here is a variant of the Euclidian distance such that it is applicable to numerical clinical variables as well as nominal clinical variables [19]. The expression for proximity is

$$d(x,y) = \sqrt{\textstyle\sum_{i=1}^{p}\emptyset_i(x_i,y_i)} \quad (2)$$

where $\emptyset_i()$ is the distance between two variables defined as follows:

$$\emptyset_i(v_1,v_2) = \begin{cases} 1 & if\ i\ is\ a\ nominal\ variable\ and\ v_1 \neq v_2 \\ 0 & if\ i\ is\ a\ nomial\ variable\ and\ v_1 = v_2 \\ (v_1-v_2)^2 & if\ i\ is\ a\ numeric\ variable \end{cases} \quad (3)$$

The ten uncensored neighbors with the smallest proximities are selected to compute the event time of interest associated with the censoring time. In addition, weights are assigned to the contributions of the neighbors, such that the nearer neighbors contribute more to the average than the more distant ones. The Gaussian function is used to obtain the weights [20]. If the neighbor is positioned at a distance d away from this censored observation, then the weight of this uncensored survival time is $w(d) = e^{-d}$.

B. Identify Latent Class Membership

In typical machine learning applications, selecting

relevant features necessitates explicit class labels being associated with the subjects, which is not possible in the survival context here. On the other hand, as will be shown in the subsequent experiments, partitioning the subjects using a straightforward "median cut" on survival time is a poor choice. In this subsection, we introduce a more sophisticated discretization method, silhouettes clustering validity [26], which we will use to partition patients into separate groups.

A prerequisite for applying the partitioning technique to discretize the continuous phenotype of interest in clinical data, which applies to the probabilistic naturalized PAM algorithm [24] used here, is the optimal choice of the number of splitting bins from the data. "The best k" of clusters to be formed should allow an appreciation of the relative quality of the clusters and overall structure of the data [25]. Unlike many discretization methods which require the number of clusters be specified in advance, silhouettes clustering validity selects the optimal number of clusters in a partitioning by evaluating the validity of the produced clustering.

Assuming the data has been clustered via a certain technique, silhouettes are constructed such that the value *s(i)*, which measures how well an object *i* has been classified, is defined for each object *i*. The computation of *s(i)* concerning a specific datum *i* in the dataset consists of calculating the average dissimilarities of *i* to the objects in the same group as *i* and those in a different group as follows:

$$s(i) = \begin{cases} 1 - {a(i)}/{b(i)} & if\ a(i) < b(i) \\ 0 & if\ a(i) = b(i) \\ {b(i)}/{a(i)} - 1 & if\ a(i) > b(i) \end{cases} \quad (4)$$

where *a(i)* is the average dissimilarities of *i* to all other objects of the cluster to which *i* has been assigned and *b(i)* holds the lowest of the average dissimilarities of *i* to the objects of the clusters in which *i* is not a member (i.e., the neighbor cluster of object *i*).

As we can see from this formula, the value of *s(i)* for each object *i* ranges from -1 to 1, and a high *a(i)* value indicates a strong dissimilarity between datum *i* and its own cluster, whereas a small *a(i)* value suggests that it is well classified within its cluster. Figure 2 shows silhouette plots of the lung cancer data for k with the four highest average silhouette widths after computing average silhouette widths for PAM partitions corresponding to all possible values of k. The silhouettes should appear as wide as possible for a natural value of k, thus one way to choose k appropriately is to select the value that yields the maximum average silhouette width. We see that the computation of average silhouette widths for all possible k returns a highest average silhouette width value of 0.72 when k = 2, so we select k = 2 to discretize the continuous time space specific to this lung cancer dataset. For an interpretation of average silhouette in determining the natural number of clusters within a dataset see [27].

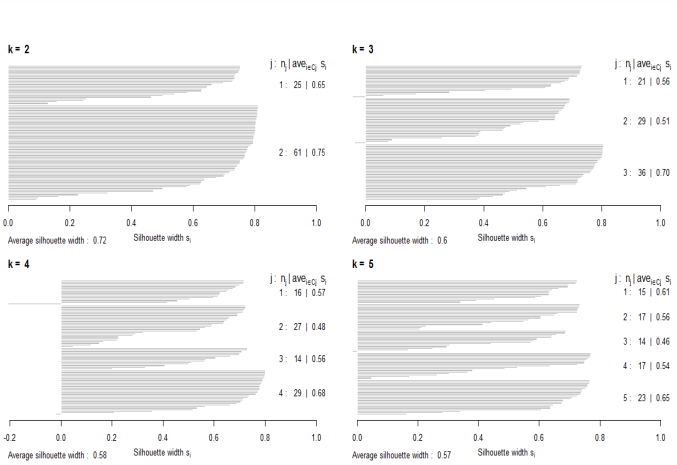

FIGURE 2 SILHOUETTES PLOTS OF THE LUNG CANCER DATA, FOR K WITH 4 TOP AVERAGE SILHOUETTE WIDTHS

C. *Feature selection by fast correlation-based filter (FCBF)*

FCBF first determines relevancy by calculating the symmetrical uncertainty (SU) for each feature related to the survival outcome, which is given by the following formula:

$$SU(X,Y) = 2\left[\frac{IG(X \mid Y)}{H(X) - H(Y)}\right] \quad (5)$$

where $H(X)$ or $H(Y)$, entropy, is a measure of the uncertainty of a random variable, and $IG(X \mid Y)$, information gain, is the amount by which the entropy of X decreases when additional information about X from Y is given.

If we have $SU(X,Y) > SU(Z,Y)$, it means Y is more correlated to X than to Z. When X and Z are genes to be selected and Y is the class label estimated from the discretization of time space, then the above formula implies that X is more relevant to Y than Z is. When SU is defined on both variables for genes, its value is used as a metric for redundancy. Two genes are considered to be redundant if the SU value defined on them exceeds one of the SU values they have with the class label. Among all the features which are redundant with each other, only the one with the highest relevance with the class label is retained.

The application of FCBF in our context is simple and efficient. A list of features (i.e., genes) is being built in descending order of their SU values, and irrelevant features are those whose SU values are less than a predefined threshold. These will be eliminated. Redundant features with lower relevancy are removed iteratively from the list until the number of remaining features reaches a targeted low bound or there are no more features to be removed. (For more detail, refer to [28].)

III. RESULTS

This section presents an empirical study designed to evaluate the performance of our proposed method on high-dimensional survival data and a comparison to related methods reported in the literature. The two real-world datasets we considered are the lung cancer dataset [5] and the renal cell dataset [14]. Before the data was half-and-half allocated for the purpose of calibration and validation, gene expressions were

excluded if their 75th percentile value was less than 100 or the variance was less than one-fifth of the interquartile range of the whole dataset. For each dataset, we run the four previously discussed statistical approaches along with our proposed procedure with two typical classifiers, Naïve Bayes (NB) and Decision Tree (DT). The leave one out cross-validation was used to determine which subtype is present in the hold out patient. Log-rank statistics comparing the survival times of different subgroups in the test cases were computed to compare the effectiveness of different methods. The efficiencies were indicated by the time needed to complete the whole leave one out trial on each dataset with respect to the selected genes in each method. These two results are shown in Table I and II, respectively.

Method	p-values	
	Lung Cancer Data	Renal Cell Data
Median-Cut	0.0285	0.054
Hierarchical Clustering	0.069	0.0461
Clustering-Cox	0.0113	0.0098
Risk Index	0.0072	0.0121
Naïve Bayes	0.0031	0.0057
Decision Tree	0.0085	0.0096

TABLE I. COMPARISON OF THE LEAVE ONE OUT APPROACH OF DIFFERENT METHODS APPLIED TO TWO DATASETS. MEDIAN-CUT, USING MEDIAN SURVIVAL TIME TO ASSIGN PATIENTS INTO CANCER SUBTYPES; HIERARCHICAL CLUSTERING, USING CLUSTERING DENDROGRAM TO ASSIGN SUBTYPES OF PATIENTS BASED ON ALL GENES; CLUSTERING-COX, USING CLUSTERING BASED ON THE GENES WITH THE LARGEST COX SCORES; RISK INDEX, USING THE CUMULATIVE EFFECTS OF THE SIGNIFICANT GENES SELECTED WITH THE LARGEST COX SCORES; NAÏVE BAYES, USING FCBF IN CONJUNCTION WITH NAÏVE BAYES CLASSIFIER; AND DECISION TREE, USING FCBF IN CONJUNCTION WITH DECISION TREE CLASSIFIER.

Method	Execution time	
	Lung Cancer Data	Renal Cell Data
Median-Cut	< 1.0	< 1.0
Hierarchical Clustering	72	102
Clustering-Cox	2241	7594
Risk Index	2417	8016
Naïve Bayes	263	1476
Decision Tree	371	1730

TABLE II. TIME TAKEN (CPU UNITS) BY DIFFERENT METHODS FOR COMPLETING A LEAVE ONE OUT TRIAL ON EACH DATASET WITH SPECIFIC TO A CERTAIN NUMBER OF SELECTED GENES.

From Table I, we can see that our method (bottom two rows) beat all other methods in statistical significance for survival prediction accuracy. In particular, Naïve Bayes in conjunction with FCBF gives the best results when 33 of the most significant genes are identified among lung cancer data and 38 genes among renal cell data, with p-values of 0.0031 and 0.0057, respectively. These results are significant predictors of survival. Table II shows the execution times reported in CPU units and verifies our proposed method merits with good scalability and is an efficient predictor of survival.

In addition to the two real-world datasets, simulation data was also used to test the effectiveness of the proposed method. We used a logistic distribution and Weibull distribution as examples to model covariates' effect on the patient's probability of being cured and the survival probability of uncured patients, respectively. The simulated clinical dataset consists of 10 covariates and 100 observations. The censoring time of each sample was generated as a uniform random number with a minimum value of 2 and a maximum value of 16. The event time was generated with a value ranging from 8 to 16 for samples 1-50, and 2 to 10 as event times for samples 51- 100. In terms of the synthetic gene expression data, the number of genes was set as 5,000, which is close to the number of genes after the preprocessing step in evaluating the lung cancer data. All gene expression values were generated as standard normal random numbers with a few exceptions: a mean of 1.0 in genes 1-50 was generated for 30% randomly selected samples, a mean of 2.0 in genes 51-200 was generated for 50% randomly selected samples, and a mean of 0.5 in genes 200-400 was generated for 70% randomly selected samples. We have now generated the clinical data and gene expression data for training. We define samples 1-50 and 51-100 as belonging to "cancer subtype 1" and "cancer subtype 2", respectively. Finally, the program ran again with exactly the same parameter settings to generate testing data. Performances of the methods discussed were compared by applying them in the identification and determination of underlying subtypes in the training and testing sets. In Table III, initial cluster errors are the number of misclassified samples when the training data is originally divided into two subgroups and prediction errors are the number of samples assigned to wrong subtypes in the testing set. For each scheme, we performed 100 runs, and took the average. From the table, we can see that the fully supervised and fully unsupervised methods performed much more poorly than other methods in this simulation study. The NB method gave the best results for both the initial cluster errors and prediction errors. The DT method was the second best in terms of initial cluster errors, losing only to NB. Its performance was slightly worse than the Risk Index method with respect to prediction errors.

Method	Initial Cluster Errors	Prediction Errors
Median-Cut	36.6	53.8
Hierarchical Clustering	37.8	37.2
Clustering-Cox	20.3	16.4
Risk Index	15.5	12.5
Naïve Bayes	7.2	4.1
Decision Tree	10.7	12.6

TABLE III. COMPARISON OF DIFFERENT METHODS APPLIED TO SIMULATION DATA.

IV. DISCUSSION

Although many studies have focused on developing statistical methods to select survival-associated features and to predict cancer subtypes from the genetic profile of a tumor, research in this area from a machine learning perspective has been underexplored. This paper incorporates and promotes the machine learning approach to predict patient survival from high-dimensional survival data. To accomplish this objective, the discretization of patient survival times via silhouettes clustering validity was used as a strategy to overcome the obstacle of hidden class information in high-dimensional survival data. The results of the real datasets and the simulation datasets conclusively match, which suggests that our proposed method is an effective and efficient predictor of survival. This work defines a new task and opens possibilities for future work on further enhancing the applications of machine learning approaches to develop more powerful diagnostic tools for cancer.

REFERENCES

[1] S. Ogino, C. S. Fuchs, and E. Giovannucci. How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. Expert Review of Molecular Diagnostics, 12(6): 621–628, 2012.

[2] M. George. Cancer: 100 different diseases. American Journal of Nursing, 66(4): 749-756, 1966.

[3] W. Choi, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer Cell, 25(2): 152–165, 2014.

[4] U. Pfeffer. Cancer Genomics: Molecular classification, prognosis and response prediction. Springer Netherlands, 2012.

[5] D. Beer, S. Kardia, C. Huang, T. Giordano, A. Levin, et al. Gene expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine 8(8): 816‐824, 2002.

[6] D. Balding. A tutorial on statistical methods for population association studies. Nature Reviews Genetics, pages 781‐791, 2006.

[7] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. PLOS Biology, 2(4): 511-522, 2004.

[8] H. Clifford, F. Wessely, S. Pendurthi and R.D. Emes. Comparison of clustering methods for investigation of genome-wide methylation array data. Frontiers in Genetics, 2: 88, 2011.

[9] A.D. Gordon. Classification. Chapman and Hall/ CRC, 1999

[10] [10] M. J. Van, Y. D. He, H. Dai, et al. A gene expression signature as a predictor of survival in breast cancer. New England Journal of Medicine, 347(25): 1999‐2009, 2002.

[11] [11] L.J van't, H. Dai, M.J. Vijver, Y.D. He, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415: 530‐536, 2002.

[12] G.D. Kleinbaum, M. Klein. Survival Analysis: A Self-Learning Text, third edition. Springer-Verlag New York, 2012.

[13] R.Tibshirani, T. Hastie, B. Narashimhan and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science, 18: 104–117, 2003.

[14] H. Zhao, B. Ljungberg, et al. Gene expression profiling predicts survival in conventional renal cell carcinoma. PLOS Medicine, 3: e13, 2006.

[15] H. Bovelstad, S. Nygard, H. Storvold, et al. Predicting survival from microarray data – a comparative study. Bioinformatics, pages 2080‐2087, 2007.

[16] X. Liu, et al. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. Statistics in Medicine, 31: 2882-2891, 2012.

[17] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling. Numerical recipes in C. Cambridge University Press, 1988.

[18] I. Choi, et al. An empirical approach to model selection through validation for censored survival data. Journal of Biomedical Informatics, 44(4): 595-606, 2011

[19] F. Mortiera, S. Robinb, S. Lassalvy, C.P. Barilc , A. Bar-Hend. Prediction of Euclidean distances with discrete and continuous outcomes. Journal of Multivariate Analysis, 97: 1799 – 1814, 2006.

[20] C.E. Rasmussen and K. I. W. Christopher. Gaussian processes for machine learning. The MIT Press, 2006.

[21] R. Tibshirani. Univariate Shrinkage in the Cox model for high-dimensional data. Statistical Applications in Genetics and Molecular Biology 8(1): Article 21, 2009.

[22] J. Lapointe, C. Li, E. Bair, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. In Proceedings of the National Academy of Sciences of the United States of America, 101(3): 811‐816, 2004.

[23] X. Cui, G. Churchill. Statistical test for differential expression in cDNA microarray experiments. Genome Biology 2003; 4: 210.

[24] A. Reynolds, et al. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modeling and Algorithms, 5: 475–504, 1992.

[25] J. Catlett. On changing continuous attributes into ordered discrete attributes. In Proceedings of the European Working Session on Learning, pages 164-178, 1991.

[26] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20: 53–65, 1987.

[27] L. Kaufman, P. J. Rousseeuw. Finding groups in data: An introduction to cluster analysis. Wiley Interscience, 2008.

[28] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 12th International Conference on Machine Learning, pages 856—863, 2003.