Feature Selection Based on Rough Set Theory Using Feature Space Decomposition for Mixed-type Data Classification

Kyung-Jun Kim
Department of Industrial and Management Engineering
Pohang University of Science & Technology (POSTECH), Pohang, Korea
Tel: (+82) 54-279-5916, Email: k_kim@postech.ac.kr

Chi-Hyuck Jun †
Department of Industrial and Management Engineering
Pohang University of Science & Technology (POSTECH), Pohang, Korea
Tel: (+82) 54-279-2197, Email: chjun@postech.ac.kr

Abstract. Feature selection plays a key role in many classification problems dealing with mixed-type data. The main idea of feature selection is to reduce the dimensionality of the input space while preserving the classification accuracy by selecting the most relevant input features. The rough set theory can be an appropriate way of measuring the importance of features in a classification problem, as seen in recent studies. Previous papers related to feature selection based on the rough set theory also considered property of mixed-type data, however, they were insufficient to investigate the properties of numerical and categorical features. To overcome the limitation, we suggest a concept of feature space decomposition and maintain the properties of each feature. In addition, for fair measure between numerical and categorical feature, we use Heterogeneous Euclidean-overlap Metric (HEOM). Finally, we conduct and show the experimental results to compare our proposed method with several benchmarking methods and select the appropriate features through the forward selection algorithm with various mixed-type data.

Keywords: Feature selection, Classification, Feature space decomposition, Mixed-type data

1. INTRODUCTION

Huge amount of features and samples are cumulated in data as time passes and this can lead to the curse of dimensionality. In this situation, feature selection is usually used to decide which features are relevant to a target feature or class label. The irrelevant features provide no useful information in any context. Furthermore, in real-world dataset, the categorical features and numerical features are usually coexisting and the great majority of feature selection algorithms are designed to work only with numerical or categorical features. There are traditional two approaches to feature selection. One is a transformation from categorical feature to numerical feature. This approach is not likely to work well and permuting the code for two categorical values could lead to different values of distance. And another is a transformation from numerical feature to categorical feature. This approach may lead to a loss of information and making the feature selection efficiency extremely depend on the discretization technique. So, we should consider the properties of mixed-type data when we conduct feature selection.

Feature selection is divided into three main methods such as filter method, wrapper method, and embedded method (Saeys *et al*., 2007). Filter methods do not incorporate learning but just directly select the best feature subset based on the intrinsic properties of the data. Filter methods are also divided into two approaches such as ranking method and space searching method (Wang *et al*., 2013). The typical examples of ranking methods are PCC (Veer *et al*., 2002), Chi-square feature selection (Liu and Stiono, 1995), Relief-F (Kononenko, 1994), Information Gain (Liu *et al*., 2002), and mRMR (Peng *et al*., 2005) and example of space searching methods is CFS (Hall, 1999). Wrapper methods use a machine learning to measure the quality of subsets of features. The typical examples of wrapper are forward selection, backward elimination, and stepwise regression (Efroymson, 1960; Kittler, 1978). Embedded methods mean the learning part and the feature selection part cannot be separated. The typical example of

embedded methods is decision tree such as CHAID (Kass, 1980), C4.5 (Quinlan, 1993), CART (Breiman *et al*., 1984) and so on.

Unlike filter method, wrapper method and embedded method use a specific classifier when they select the best feature subset, so the result of these methods depends on kinds of classifier. In addition, they have high time complexity.

In this paper, we will introduce a rough set theory for feature selection and literature reviews of related papers briefly. And then, we will propose improved approach by using feature space decomposition and Heterogeneous Euclidean-overlap Metric (HEOM) for mixed-type data. The rest of the paper is organized as follows: literature reviews of related papers are present in Section 2. Section 3 gives improved propose methods and Section 4 gives description of dataset and experimental results. Conclusion is in Section 5.

2. A REVIEW ON ROUGH SET THEORY

In this section, we consider literature review of preview papers briefly related to feature selection based on rough set theory for mixed-type data.

2.1 Rough Set Theory

In Pawlak's rough set model (Pawlak, 1991), the objects with the same feature values in terms of features B are drawn together and form an equivalence class, denoted by $[x]_B$. Equivalence classes are also called elemental information granules or elemental concepts. The family of elemental granules $\{[x_i]_B, x_i \in U\}$ builds a concept system to describe arbitrary subset of the sample space, where U denotes total sample set. Then two unions of elemental granules are associated with subset $X \subset U$: lower approximation and upper approximation

$$\underline{B}X = \{[x_i]_B | [x_i]_B \subseteq X\} \quad (1)$$

$$\overline{B}X = \{[x_i]_B | [x_i]_B \cap X \neq \emptyset\} \quad (2)$$

The lower approximation is the maximal union of elemental granules consistently contained in X and the upper approximation is the minimal union of elemental granules containing X. The difference between lower approximation and upper approximation is called approximation boundary of X and the equation defined as

$$BN(X) = \overline{B}X - \underline{B}X \quad (3)$$

Rough set based feature selection is to find minimal subset of feature and the decision has maximal consistent elemental granules in terms of the selected features.

2.2 Rough Set Theory for Numerical Feature

For categorical features, we can just use the Pawlak's rough set model as described in section 2.1. For numerical features, however, we cannot use the Pawlak's rough set model directly. Hence, there is also an approach to deal with numerical features as below Definition 1 and Definition 2 by (Pawlak, 1991).

Definition 1. Given a set of finite and nonempty objects $U = \{x_1, x_2, \dots, x_n\}$ and a numerical attribute a to described the objects, the δ neighborhood of arbitrary object $x_i \in U$ is defined as

$$\delta_a(x_i) = \{x_j | \Delta(x_i, x_j) \leq \delta, x_j \in U\}, \text{ where } \delta \geq 0 \quad (4)$$

where $\Delta(x_i, x_j)$ denotes a metric and the most frequently used metric is Euclidean distance.

Definition 2. Given arbitrary subset X of the sample space and a family of neighborhood information granules $\delta_a(x_i)$, $i = 1, 2, \dots, n$, we define the lower and upper approximations of X with respect to neighborhood relation R_a as

$$\underline{R_a}X = \{x_i | \delta(x_i) \subseteq X, x_i \in U\} \quad (5)$$

$$\overline{R_a}X = \{x_i | \delta(x_i) \cap X \neq \emptyset, x_i \in U\} \quad (6)$$

The difference of lower and upper approximations is called the boundary of X:

$$BN(X) = \overline{R_a}X - \underline{R_a}X \quad (7)$$

2.3 Rough Set Theory for Mixed-type Data

A neighborhood information system is denoted by $NIS =< U, A, V, f >$, where U is the sample set, called the universe, A is the attribute set, V is the domain of attribute values. f is an information function $f: U \times A \rightarrow V$. More specifically, a neighborhood information system is also called a neighborhood information system is also called a neighborhood decision table if there are two kinds of attributes in the system: condition and decision, which is denoted by $NDT =< U, A \cup D, V, f >$. Given $NIS =<$

$U, A, V, f >$, $B = B^n \cup B^c$, where B^n and B^c are subsets of numerical features and categorical features, respectively, B^n generates neighborhood relation R_{B^n} and B^c generates equivalence relation R_{B^c}, the neighborhood granule of x in terms of features B is defined as

$$R_B(x) = \{x_i | x_i \in R_{B^n}(x) \wedge x_i \in R_{B^c}(x), \forall a_i \in B^n, b_j \in B^c\} \quad (8)$$

Also, given a neighborhood decision table $NDT =< U, A \cup D, V, f >$, $X_1, X_2, \ldots, X_N$ are the subsets of objects with decisions 1 to N, $R_B(x_i)$ is the neighborhood information granules including x_i and generated with mixed features $B \subseteq A$, then the lower approximation of decision D with respect to B is defined as

$$\underline{R_B} D = \left\{ \underline{R_B} X_1, \underline{R_B} X_2, \ \ldots, \underline{R_B} X_N \right\} \quad (9)$$

The dependency degree of D to B is defined as the ratio of consistent objects:

$$\gamma_B(D) = \frac{|POS_B(D)|}{|U|} \quad (10)$$

where $POS_B(D)$ denotes the lower approximation of decision.

Dependency function reflects the describing capability of features B, which can be considered as the significance of features B to approximate decision D. Through this value, the significance of feature a relative to B and D is defined as

$$SIG(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D) \quad (11)$$

By using $SIG(a.B, D)$, features in mixed-type data could be selected through feature selection algorithms. For more detailed information about the rough set theory for mixed-type data, the readers can refer to (Hu *et al*., 2008; He *et al*., 2010).

3. PROPOSED METHOD

This section describes the proposed method called RST_FSD (Rough Set Theory using Feature Space Decomposition) in detail. Figure 1 illustrates the overall procedure of RST_FSD, where U_X denotes the candidate numerical feature set, U_Z denotes the candidate categorical feature set, and S denotes selected feature set. For the very first step, numerical and categorical features are ordered by ranking method separately. The numerical features are ranked by ERGS (Chandra and Gupta, 2011) which has a similar property with rough set and the categorical features are ranked by equivalence relation which can be directly induced from categorical features based on the feature values. After that, feature space decomposition and Heterogeneous Euclidean-overlap metric (HEOM) (Wilson and Martinez, 1997) are applied to search strategy. Now, ERGS, feature space decomposition, and HEOM are described briefly.

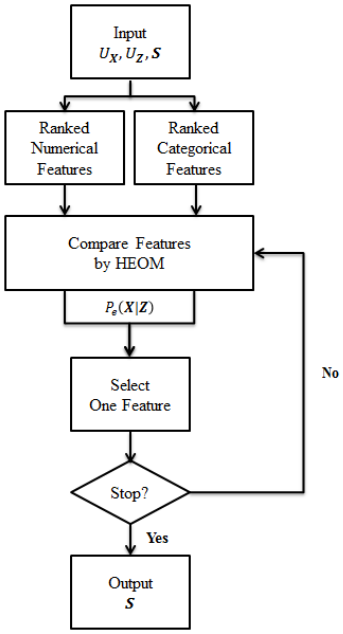

Figure 1: Procedure of RST_FSD

3.1 Rough Set Theory for Numerical Feature

Suppose that a dataset consists of d input features $\{a_1, a_2, \ldots, a_d\}$ and a single response representing the class label from 1 to l. Suppose also that there are N observations on these features and the response. Let μ_{ij} and σ_{ij} denote the mean and standard deviation of the i^{th} feature for class j, respectively. The effective range of i^{th} feature a_i for the j^{th} class, denoted by R_{ij}, is defined by the range between the lower bound, r_{ij}^-, and the upper bound, r_{ij}^+,

$$R_{ij} = \left[r_{ij}^-, \ r_{ij}^+ \right] \quad (12)$$

where the lower and upper bounds are obtained by

$$r_{ij}^- = \mu_{ij} - (1 - p_j)\gamma\sigma_{ij} \quad (13)$$

$$r_{ij}^+ = \mu_{ij} + (1 - p_j)\gamma\sigma_{ij} \quad (14)$$

The prior probability of jth class is p_j. Here, the

factor $(1-p_j)$ is taken to scale down effect of class with high probabilities and consequently large variance. The value of γ is determined statistically by Chebyshev inequality defined as

$$P\left(\left|X-\mu_{ij}\right| \geq \gamma\sigma_{ij}\right) \leq \frac{1}{\gamma^2} \quad (15)$$

which is true for all distributions. The value of γ is set as 1.732 for the effective range which contains at least 2/3rd of the data objects.

The overlapping area (OA_i) among the classes of the i^{th} feature is defined as

$$OA_i = \sum_{j=1}^{l-1}\sum_{k=j+1}^{l}\varphi_i(j,\ k) \quad (16)$$

where

$$\varphi_i(j,\ k) = \begin{cases} r_{ij}^{+} - r_{ik}^{-} & if\ r_{ij}^{+} > r_{ik}^{-} \\ 0 & otherwise \end{cases} \quad (17)$$

By scaling and normalizing *OA*, the weights for features can be calculated using

$$w_i = 1 - AC_i/AC_{max} \quad (18)$$

where

$$AC_i = \frac{OA_i}{\max_j\left(r_{ij}^{+}\right) - \min_j\left(r_{ij}^{-}\right)} \text{ and } AC_{max} = \max(AC_i) \quad (19)$$

Through the w_i, features could be ordered as descending order. For more detailed information about the ERGS algorithm, the readers can refer to (Chandra and Gupta, 2011).

3.2 Feature Space Decomposition

In the feature space defined by a mixed feature set with both numerical and categorical features, that is $[\boldsymbol{X}, \boldsymbol{Z}]$, the error probability, $P_e(\boldsymbol{X}, \boldsymbol{Z})$, can be written with class $y_j, j = 1, 2, \ldots, C$ as

$$P_e(\boldsymbol{X}, \boldsymbol{Z}) = \sum_{\boldsymbol{Z}}\int_{\boldsymbol{X}}\left[1 - \max_j p\left(y_j|\boldsymbol{X}, \boldsymbol{Z}\right)\right]p(\boldsymbol{X}, \boldsymbol{Z})\, d\boldsymbol{X} \quad (20)$$

Rearranging (20), yields as

$$P_e(\boldsymbol{X}, \boldsymbol{Z}) = 1 - \sum_{\boldsymbol{Z}}\int_{\boldsymbol{X}}\max_j p\left(y_j, \boldsymbol{X}, \boldsymbol{Z}\right) d\boldsymbol{X}$$

$$= \sum_{Z} p(\boldsymbol{Z}) - \sum_{\boldsymbol{Z}}\int_{\boldsymbol{X}}\max_j p\left(y_j, \boldsymbol{X}|\boldsymbol{Z}\right) p(\boldsymbol{Z})\, d\boldsymbol{X}$$

$$= \sum_{Z} p(\boldsymbol{Z})\int_{\boldsymbol{X}}\left[1 - \max_j p\left(y_j|\boldsymbol{X}, \boldsymbol{Z}\right)\right] p(\boldsymbol{X})\, d\boldsymbol{X}$$

$$= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}) P_e(\boldsymbol{X}|\boldsymbol{Z}) = \sum_{i=1}^{N} p(z_i) P_e(\boldsymbol{X}|z_i) \quad (21)$$

where the maximum likelihood estimate of $p(z_i)$ is the ratio of the frequency of samples given the value-combination z_i, denoted as n_i, to the number of samples n, that is $\hat{p}(z_i) = n_i/n$. If there are selected categorical features, the feature space is decomposed into a set based on the multi-nominal feature $\boldsymbol{z}$. Also $\hat{P}_e(\boldsymbol{X}|z_i)$ can be calculated as

$$\hat{P}_e(\boldsymbol{X}|z_i) = 1 - \frac{1}{n_i}\sum_{l=1}^{n_i}\frac{r(l)^{-d}}{\sum_{j=1}^{C} r_j(l)^{-d}} \quad (22)$$

where $r(l)^{-d}$ is the volume which is centered at x_l in d-dimensional space defined by the features $\boldsymbol{X}$.

3.3 Heterogeneous Euclidean-overlap Matric

The HEOM is defined as

$$HEOM(x, y) = \sqrt{\sum_{a=1}^{m} d_a(x_a, y_a)^2} \quad (23)$$

where m is the number of features and $d_a(x_a, y_a)$ is the distance between samples x and y in terms of feature a, which is defined as

$$d_a(x_a, y_a) = \begin{cases} 1, & \text{if the feature value of } x \text{ or } y \text{ is unknown} \\ overlap_a(x, y), & \text{if } a \text{ is a categorical feature} \\ rn\_diff_a(x, y), & \text{if } a \text{ is a numerical feature} \end{cases} \quad (24)$$

Here

$$overlap_a(x, y) = \begin{cases} 0, & if\ x = y \\ 1, & otherwise \end{cases} \quad (25)$$

$$rn\_diff\_a\ (x, y) = |x - y|/(max_a - min_a) \quad (26)$$

3.4 Feature Selection Based on RST_FSD

At the first step, the numerical and categorical features are ordered as descending order by ranking method separately. The numerical features are ranked by ERGS (Chandra and Gupta, 2011) which has a similar property with rough set and the categorical features are ranked by equivalence relation which can be directly induced from categorical features based on the feature values.

At the second step, the highest ranked candidate numerical feature and categorical feature are chosen and calculate the rough set error using HEOM and δ is an user specific parameter which is for controlling the granularity of a neighborhood approximation space. The rough set error is calculated as

$$Error = \frac{misclassified\ \ samples}{total\ \ samples} \quad (27)$$

Then, select one of two candidate features that the error is lower and save in selected feature set S if the error is lower than previous step error. If not, the selected candidate feature is removed from candidate feature set.

At the third step, if saved feature at the second step is categorical feature, then feature space decomposition is conducted through the values of that feature. If saved feature is numerical feature, then there is no decomposition, but it is used with the candidate numerical feature and categorical feature when we compare the errors using HEOM at second step.

Although selected candidate feature is categorical feature, if there is very small number of samples in decomposed space, there is no more feature space decomposition. Hence, the candidate categorical feature is just used with selected numerical features by HEOM.

All of above steps are conducted iteratively until there is no candidate feature and selected feature set S can be obtained.

Table 1: Summary of the Datasets

Dataset	Observations	Numerical	Categorical	Classes
Credit Approval	690	6	9	2
German	1,000	7	13	2
Heart Disease	303	6	7	2
Heart Statlog	270	6	7	2
Hepatitis	155	6	13	2
Horse-colic	368	7	15	2
Housing	506	12	1	2

Table 2: Number of Selected Features (N: Numerical, C: Categorical, T: Total)

Dataset	Raw data			Relief-F			Information Gain			CFS			NDEM			RST_FSD		
	N	C	T	N	C	T	N	C	T	N	C	T	N	C	T	N	C	T
Credit Approval	6	9	15	0	7	7	4	3	7	4	3	7	4	2	6	3	6	9
German	7	13	20	3	5	8	3	5	8	1	4	5	4	8	12	3	3	6
Heat Disease	6	7	13	1	6	7	3	4	7	3	4	7	6	6	12	1	2	3
Heart Statlog	6	7	13	1	6	7	3	4	7	3	4	7	6	6	12	1	2	3
Hepatitis	6	13	19	0	8	8	2	6	8	2	5	7	2	9	11	1	4	5
Horse-colic	7	15	22	1	6	7	1	6	7	1	4	5	4	3	7	1	7	8

Housing	12	1	13	7	0	7	7	0	7	4	0	4	8	0	8	3	1	4
Average	16.43			7.29			7.29			6.00			9.71			**5.43**		

Table 3: Classification Performances of Seven Mixed-type Datasets with RBF-SVM (Acc: Accuracy, Std: Standard Deviation)

Dataset	**Relief-F**		**Information Gain**		**CFS**		**NDEM**		**RST_FSD**	
	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std
Credit Approval	86.42	0.4873	86.37	0.4131	86.37	0.3553	86.37	0.4431	**87.29**	**0.4560**
German	73.84	0.5211	75.25	0.2677	75.19	0.3281	75.29	0.3725	**75.65**	**0.6932**
Heart Disease	82.93	1.2391	83.48	0.3717	83.74	0.9151	82.41	0.7857	**85.07**	**0.1789**
Heart Statlog	83.78	0.6717	83.85	1.0060	**84.63**	**0.6591**	83.69	0.8083	83.87	0.6053
Hepatitis	83.75	0.8355	88.38	1.6719	92.88	1.3242	87.88	1.2430	**97.38**	**0.3953**
Horse-colic	86.44	0.6162	85.82	0.4074	83.91	0.8364	**86.90**	**0.4600**	85.61	0.6708
Housing	93.46	0.6676	92.86	0.4022	94.57	0.2522	92.09	0.5869	**97.79**	**0.2909**
Average	84.37		85.14		85.90		84.95		**87.52**	

4. EXPERIMENTAL RESULT

4.1 Dataset

For the experiments, 7 datasets are used from UCI Machine Learning Repository. These are mixed-type dataset which contain both numerical and categorical features. The datasets are summarized in Table 1.

4.2 Experiment Setting

In this paper, we choose the four feature selection methods; Relief-F, Information Gain, CFS, NDEM (Hu *et al.,* 2010), and RST_FSD which the output of the algorithm is a subset of features, not ranking except for Relief-F and Information Gain. We select top k features from Relief-F and Information Gain where k denotes mean number of CFS, NDEM and RST_FSD.

We use RBF-SVM as classifier and experiments are conducted by 5 cross-validations with 10 repetitions to get accuracies. We set the values of δ =0.005, 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18, 0.20, 0.25, 0.30, and 0.40. And there is no decomposition if the number of samples in subset is less than 25. If there are missing values in dataset, we delete that observation. The Horse-colic, however, contains a lot of missing values, we conduct data imputation which mean for numerical feature and mode for categorical feature.

4.3 Results and Discussion

Table 2 and Table 3 show the performances of five feature selection methods. As shown in Table 3, our proposed method RST_FSD shows better performances than the other methods on the five of seven datasets such as Credit Approval, German, Heart Disease, Hepatitis, and Housing. Especially, RST_FSD outperforms the other methods on Hepatitis and Housing. Also, the total average of classification performances of RST_FSD is the highest on these seven datasets.

Actually, the differences of performance on Credit Approval, German, Heart Statlog, and Horse-colic are so marginal. As shown in Table 2, however, RST_FSD can make similar performances with only a relatively small number of selected features.

In addition, on the Housing dataset, Relief-F, Information Gain, CFS, and NDEM cannot capture any categorical feature. But, RST_FSD can select categorical feature because our proposed method is based on intuitive feature space decomposition.

5. Conclusion

In this paper, we proposed improved feature selection method based on rough set theory and compare with Relief-F, Information Gain, CFS, and NDEM. Our method has the differentiation factor such as HEOM and feature space decomposition.

For future work, we will conduct more experiments through other classifiers such as Naïve Bayes and CART or other benchmark feature selection methods. In addition, we

will collect multi-class datasets and do the same procedure on that datasets.

ACKNOWLEDGMENTS

Thins research was supported by National Research Foundation of Korea
(Project No. 2013R1A2A2A03068323)

REFERENCES

Blake, C., and Merz, C. J. (1998). {UCI} *Repository of machine learning databases*. http://www.ics.uci.edu/~mlearn/MLRepository.html

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A., (1984). *Classification and regression trees*. Wadsworth and Brooks, Monterrey, CA.

Chandra, B., Gupta, M., (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics*, **44**(4), 529-535.

Efroymson, M., (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, **1**, 191-203.

Hall, M.A., (1999). Correlation-based feature selection for machine learning. (Doctoral dissertation, The University of Waikato).

Hu, Q., Liu, J., and Yu, D. (2008). Mixed feature selection based on granulation and approximation. *Knowledge-Based Systems*, **21**(4), 294-304.

Hu, Q., Pedrycz, W., Yu, D., and Lang, J. (2010). Selecting discrete and continuous features based on neighborhood decision error minimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **40**(1), 137-150.

Kass, G.V., (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**(2), 119-127.

Kittler, J., (1978). Feature set search algorithms. *Pattern Recognition and Signal Processing* (eds. Sijthoff and *ence research*, **6**, 1-34.

Noordhoff). Alphen aan den Rijn, 41-60.

Kononenko, I. (1994, April). Estimating attributes: analysis and extensions of RELIEF. *In Machine Learning: ECML-94* (pp. 171-182). Springer Berlin Heidelberg.

Liu, H., Li, J., and Wong, L., 2002. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, **13**, 51-60.

Liu, H., and Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. *Paper presented at the 2012 IEEE 24th International Conference on Tools with Artificial Intelligence*.

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data Kluwer Academic Publishers. Dordrecht*, The Netherlands.

Peng, H., Long, F., and Ding, C., (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(8), 1226-1238.

Quinlan, J.R., (1993). *C4.5: Programs for Machine Learning* (Vol.1). Morgan Kaufmann, San Francisco, CA

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, **23**(19), 2507-2517.

Van't Veer, L.J., Dai, H., Van De Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., and Witteveen, A.T., (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530-536.

Wang, J., Wu, L., Kong, J., Li, Y., and Zhang, B. (2013). Maximum weight and minimum redundancy: a novel framework for feature subset selection. *Pattern Recognition*, **46**(6), 1616-1627.

Wilson, D. R., and Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of artificial intellig*