A Feature Selection Framework Based on Supervised Data Clustering

Hongzhi Liu, Bin Fu, Zhengshen Jiang, Zhonghai Wu
School of Software and Microelectronics,
Peking University, Beijing, 102600, P. R. China
Email: {liuhz, fubin, jiangzhengshen, wuzh}@pku.edu.cn

D. Frank Hsu
Department of Computer and Information Science,
Fordham University, New York, NY, 10023, USA
Email: hsu@cis.fordham.edu

Abstract—**Feature selection is an important step for data mining and machine learning to deal with the curse of dimensionality. In this paper, we propose a novel feature selection framework based on supervised data clustering. Instead of assuming there only exists low-order dependencies between features and the target variable, the proposed method directly estimates the high-dimensional mutual information between a candidate feature subset and the target variable through supervised data clustering. In addition, it can automatically determine the number of features to be selected instead of manually setting it in a prior. Experimental results show that the proposed method performs similar or better compared with state-of-the-art feature selection methods.**

Keywords—feature selection; supervised data clustering; mutual information; classification

I. Introduction

In the era of big data, it is easier to collect more information and extract more features for various applications. However, too many features will result in the curse of dimensionality, i.e. increasing the computational cost of learning and degrading the performance of prediction. Feature selection plays a key role in this situation. It is a process to find the smallest possible relevant feature subset while discarding irrelevant and redundant features [1]. It can be used to improve the prediction performance, provide faster and more cost-effective predictors, and provide a better understanding of the prediction results [2]. Because of these benefits, feature selection has been used in various applications, including information retrieval, image classification, and bio-medical data analysis [1].

However, feature selection is an NP-hard problem. The number of candidate feature subsets increases exponentially as the size of feature set increases. The sizes of the optimal feature subset for different data sets are always different. This makes it difficult to determine the optimal parameter settings for feature selection algorithms that require setting the size of feature subset to be selected in a prior. In this paper, we propose a feature selection framework which can automatically determine the number of features to be selected.

Feature selection methods could be roughly grouped into three categories: wrapper, embedded, and filter [3]. Wrapper methods take the prediction performance of a specific predictor with the candidate feature subsets as scores, and select the feature subset with the highest score. Embedded methods perform feature selection during the construction of predictors, e.g. the feature selection at each node of a decision tree. Filter methods perform as a preprocessing step which is based on an evaluation metric calculated directly from the data, without feedback from predictors. Compared with wrapper methods and embedded methods, which are specific to a chosen predictor, filter methods provide a generic selection of features, not tuned for/by a given predictor.

Mutual information is a commonly used metric for filter feature selection [4, 5]. It is a nonparametric and nonlinear measure of relevance between two variables. It does not rely on any predictors, but provide a bound on the error rate using any predictor for a given distribution [6, 7]. However, estimation of high-dimensional mutual information is still a challenging problem [8]. To avoid calculating the high-dimensional mutual information, many researchers propose to use some low-dimensional approximation methods [9, 10, 11]. They assume that there are only low-order dependencies between features and the target variable, which is not true for most real data sets.

In this paper, we propose a novel feature selection framework based on data clustering. It directly estimates the high-dimensional mutual information between a feature subset and the target variable through supervised data clustering.

The rest of this paper is organized as follows. The related work on feature selection is provided in Section II. Section III describes the proposed framework and some algorithms in detail. Empirical evaluation of the proposed method and comparison with other state-of-the-art feature selection algorithms are presented in Section IV. Section V concludes the paper.

II. Related Work

The literature on feature selection is rich. The feature selection methods can be categorized from different perspectives: 1) supervised versus unsupervised, 2) filter versus embedded versus wrapper, 3) univariate versus multivariate, 5) global versus local, 6) direct versus incremental; 4) forward versus backward, 7) evaluation criterion used for comparing candidate feature subsets.

Supervised (class-aware or class dependent) feature selection methods consider the class information when make a decision whereas unsupervised (class-blind or class

This work is partially sponsored by National Natural Science Fund of China (GrantNo.61232005).

independent) ones do not. Global feature selection methods make a decision by taking into account all data in a context-free way while local methods only use some local or context dependence information of the data. Direct feature selection methods select a subset of features simultaneously. In contrast, incremental methods begin with a simple subset (empty or the whole) and then pass it through an improvement process. According to the different procedures used to update the subset, incremental methods are divided into two sub-categories: forward increment and backward elimination. Feature selection with forward increment starts with an empty subset and then iteratively selects the best one in the retained feature set and moves it into the selected feature subset. In contrast, feature selection with backward elimination starts with the whole feature set and then iteratively remove the least important one from this set. Univariate feature selection methods assume the features are independent with each other and select the top *k* features as the result according to some criterion. In contrast, multivariate feature selection methods take into account the interactions between features during selection. According to the moment and interaction related with the learner, we can classify a feature selection method as filter, embedded or wrapper. Filter feature selection act as a preprocessing step for learners and they are independent from the learning algorithms. Embedded methods select features during the construction of classifiers and themselves are a subpart of the classifiers. Wrapper methods rely on the accuracy of a selected classifier. The relevance and redundancy of one or a subset of features can be measured using different criteria. Several commonly used criteria include correlation, distance between distribution, measure based on information theory, consistency, and classification accuracy. More detailed discussions and comparisons about different feature selection methods can be found in [1] and [3].

Song et al. [12] proposed a feature selection based on clustering. They first divided features into clusters through minimum spanning tree generation which is a graph-theoretic clustering method. Then they selected the most representative feature from each cluster that is strongly related to target classes to form a subset of features. Different from this method which uses graph clustering to remove the redundant features, our method uses data clustering to estimate the mutual information between a feature subset and the target variable.

III. FEATUE SELECTION BASED ON SUPERVISED DATA CLUSTERING

Let $\boldsymbol{F}$ be a set of features and *C* be the class label. The goal of feature selection is to find the smallest feature subset $\boldsymbol{F}_S \subset \boldsymbol{F}$ that maximizes the mutual information between the feature subset $\boldsymbol{F}_s$ and the class variable *C*. According to the Hellman-Ravi inequality [13], mutual information provides an upper bound for the Bayes error, which is the ultimate criterion (golden standard) for any procedure related to discrimination, i.e.

$$e_{Bayes} \leq \frac{1}{2}(H(C) - I(\boldsymbol{F}_S; C)) .$$

As the entropy of the class variable $H(C)$ is fixed when the data set is given, the larger the mutual information between the selected feature subset and the class variable $I(\boldsymbol{F}_S; C)$, the lower of the Bayes error's upper bound.

The main difficulty is to estimate the high-dimensional mutual information between a feature subset and the class variable $I(\boldsymbol{F}_S; C)$. We solve this problem through data clustering, i.e.

$$I(\boldsymbol{F}_S; C) \approx \max_g I(g(\boldsymbol{F}_S); C)$$

where $g(.)$ is a clustering operation which transforms the high-dimensional feature space $\boldsymbol{F}_s$ into an one-dimensional cluster label space.

A. Estimation of Mutual Information Using Supervised Data Clustering

To estimate the high-dimensional mutual information between a feature set $\boldsymbol{F}_s$ and the class variable *C*, we proposed a Supervised data Clustering Framework based on Mutual Information (SCFMI). *The main idea of this framework is iteratively increasing the number of clusters until the mutual information between the cluster label and the class label does not increase any more. Therefore, it can automatically determine the optimal number of clusters without manual intervention and output the mutual information between the feature set and the class variable as a by-product.*

When combined with different clustering methods, we can get different implementations of this framework. The following is the pseudo-code of supervised hierarchical clustering algorithm based on mutual information (SCFMI-HC), which is an implementation of the proposed framework using hierarchical clustering. It outputs the maximal mutual information between a feature set and the class variable estimated by supervised hierarchical clustering. First, it constructs a binary tree from bottom to top by successively merging similar groups of points (Fig.1). Then, it iteratively tries to divide the data into different clusters by cutting the tree at different levels from top to down, and calculates the mutual information (MI) between the cluster label and the class label. The iteration process stops when it cannot increase the MI anymore by further dividing the clusters.

Algorithm: SCFMI-HC

```
1. Initialization: maxMI←0
2. Construct a binary clustering tree from bottom to top by
   successively merging similar groups of points
3. Iteratively cut the clustering tree from top to down
   3.1 Calculate MI between cluster label and class label
   3.2 If MI>maxMI
           maxMI←MI
       Else
           Stop the iteration
4. Output the maxMI.
```

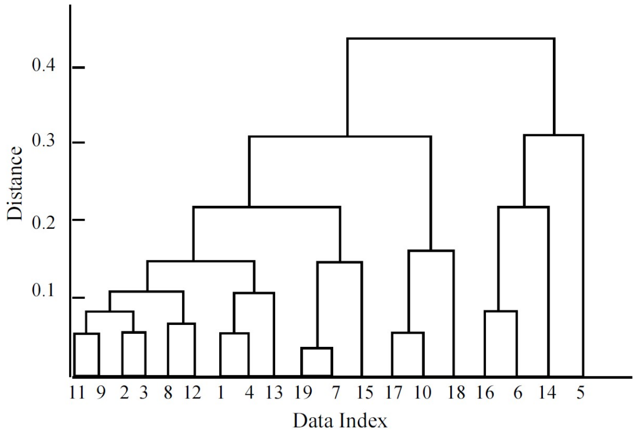

Fig. 1. A clustering tree constructed by successively merging similar groups of points

We can design other SCFMI algorithms to estimate the mutual information using different clustering methods such as density peak based clustering [14], k-means [15], and self-organizing map based clustering [16] in a similar way as SCFMI-HC.

B. Towards Optimal Feature Selection via SCFMI

The proposed feature selection algorithm FSDC (Feature Selection based on Data Clustering) consists of two main steps. First, the selected feature set $\boldsymbol{F}_s$ is initialized to the empty set and the candidate feature set $\boldsymbol{F}$ is initialized to the whole feature set. Then, it repeatedly searches the feature f_k from $\boldsymbol{F}$ that maximizes the increment of mutual information between the selected feature set $\boldsymbol{F}_s$ and the class variable C by adding f_k into $\boldsymbol{F}_s$, until it cannot find any feature that increases the mutual information between the selected feature set $\boldsymbol{F}_s$ and the class variable C. The mutual information between a candidate feature subset and the class variable is estimated via a SCFMI algorithm. In each iteration, we remove the found feature f_k from $\boldsymbol{F}$ and add it into $\boldsymbol{F}_s$.

Algorithm: FSDC

```
1. Set $\boldsymbol{F}$ to be the whole feature set and $\boldsymbol{F}_s$ to be the empty set
2. Iteratively move a feature from $\boldsymbol{F}$ to $\boldsymbol{F}_s$
   2.1 Calculate $I(C; \boldsymbol{F}_s \cup \{f_i\})$ for all $f_i \in \boldsymbol{F}$ via SCFMI
   2.2 Choose $f_k \in \boldsymbol{F}$ that maximizes $I(C; \boldsymbol{F}_s \cup \{f_i\})$
   2.3 If $I(C; \boldsymbol{F}_s \cup \{f_i\}) > I(C; \boldsymbol{F}_s)$
           $\boldsymbol{F}_s \leftarrow \boldsymbol{F}_s \cup \{f_i\}$
           $\boldsymbol{F} \leftarrow \boldsymbol{F} - \{f_i\}$
       Else
           Stop the iteration
3. Output the selected features set $\boldsymbol{F}_S$.
```

FSDC is a forward filter feature selection method based on mutual information. Two key components of this kind of feature selection methods are the calculation of the joint mutual information $I(C; \boldsymbol{F}_s \cup f_i)$ and the criterion used to stop the searching. Different from previous methods, FSDC uses a novel algorithm based on data clustering to estimate the joint mutual information. Instead of using the number of selected features as the stop criterion, FSDC directly uses the joint mutual information as the stop criterion, i.e. stop searching when it cannot find any feature to increase the mutual information between the selected feature set and the class variable.

IV. EXPERIMENTS

The performance of the proposed feature selection algorithms are evaluated and compared with several state-of-the-art feature selection methods.

A. Experimental Setup

Data sets:

Ten public benchmark data sets from UCI machine learning repository [17] and NIPS Feature Selection Challenge are used as experimental data. Table I gives a summary of the data sets.

TABLE I. DESCRIPTION OF EXPERIMENTAL DATA SETS

Data set	# of Samples	# of Features	# of Classes
Breast	569	30	2
Congress	435	16	2
CTG	2126	21	3
Ionosphere	351	34	2
Krvskp	3196	36	2
MADELON	2600	500	2
PageBlock	5472	10	5
PenBased	10992	16	10
Texture	5500	40	11
Waveform	5000	40	3

Classifiers:

To evaluate the performance of feature selection algorithms, we feed the data sets before and after feature selection into three well-known classifiers: k-Nearest Neighbor (kNN), Naïve Bayes classifier (NB) and C4.5, and record their classification accuracies. KNN is a type of instance-based lazy learning: an object is classified by a majority vote of its *k* nearest neighbors. It is one of the simplest and most effective machine learning methods. A naïve Bayes classifier is a simple probabilistic classifier based on the Bayesian theorem with strong (naïve) independence assumptions. C4.5 is a state-of-the-art method for inducing decision trees using the concept of information entropy.

Ten-fold cross-validation is adopted to better evaluate the performance of classifiers and feature selection algorithms. Each data set is randomly divided into ten equally sized subparts. Nine of the ten subparts are used as training data and the remaining one is used as testing data. The cross-validation process is repeated ten times with each of the ten subparts used exactly once as testing data. Except the MADELON data set from NIPS Feature Selection Challenge, which has been pre-divided into a training subset and a validation subset, all the experimental results are recorded as the average of the ten runs of 10-fold cross-validation.

Clustering methods:

Besides hierarchical clustering, we select three other representative clustering algorithms to implement the proposed feature selection framework, including density peak based clustering [14], k-means[15], and self-organizing map based clustering [16]. The corresponding feature selection methods are denoted as FSDC(DP), FSDC(kMeans), and FSDC(SOM) respectively.

B. Experiment 1:

The goal of this experiment is evaluate the effects of the proposed feature selection methods, i.e. comparing the performance of different classifiers with and without using the proposed feature selection methods.

Experimental design:

Using the ten benchmark data sets, we evaluate the performance of feature selection methods by the classification accuracy of different classifiers, including kNN, naïve Bayes, and C4.5.

The classification accuracy on the data with only fewer selected features is expected to be similar or higher as compared to that on the original data with all features. To evaluate the effects of feature selection, we compare the accuracy of classifiers on both the original data sets without feature selection and the data sets after feature selection.

Experimental results:

Table II shows the number of feature selected by different methods. Table III, Table IV, and Table V show the classification accuracies of kNN, Naïve Bayes, and C4.5 with alternative feature selection methods, respectively. The results on the original data sets are used for comparison.

TABLE II. NUMBER OF FEATURES SELECTED BY DIFFERENT METHODS

Data set	**Orig.**	**FSDC (DP)**	**FSDC (HC)**	**FSDC (SOM)**	**FSDC (kMeans)**
Breast	30	4.3	3.9	3.8	5.1
Congress	16	3.3	5.8	3.2	3.1
CTG	21	3.7	5.6	5.2	6.2
Ionosphere	34	3.2	4.1	3.0	3.9
Krvskp	36	5.9	12.6	6.2	7.0
MADELON	500	3.0	4.0	7.0	8.0
PageBlock	10	3.0	3.1	3.9	3.9
Penbased	16	9.1	9.6	10.3	9.1
Texture	40	7.0	5.0	6.7	6.1
Waveform	40	3.2	12.6	7.7	11.5
Avg.	74.3	**4.6**	6.6	5.7	6.4

TABLE III. CLASSIFICATION ACCURACY OF KNN (K=3) WITH AND WITHOUT FEATURE SELECTION (%)

Data set	**Orig.**	**FSDC (DP)**	**FSDC (HC)**	**FSDC (SOM)**	**FSDC (kMeans)**
Breast	96.8	95.1	93.5	93.5	95.1
Congress	91.5	94.0	96.1	93.8	95.9
CTG	90.8	88.1	89.2	91.3	88.4
Ionosphere	86.3	90.0	88.3	91.7	87.7
Krvskp	96.4	89.7	94.0	97.3	94.2
MADELON	49.7	73.0	80.3	90.7	91.0
PageBlock	96.0	96.0	95.7	96.1	96.2
Penbased	99.3	95.8	98.0	98.2	96.6
Texture	98.7	94.5	96.1	93.9	95.0
Waveform	76.3	70.5	75.6	79.6	78.7
Avg.	88.2	88.7	90.7	**92.6**	91.9

TABLE IV. CLASSIFICATION ACCURACY OF NAÏVE BAYES WITH AND WITHOUT FEATURE SELECTION (%)

Data set	**Orig.**	**FSDC (DP)**	**FSDC (HC)**	**FSDC (SOM)**	**FSDC (kMeans)**
Breast	93.0	94.6	93.5	94.0	94.9
Congress	91.7	94.9	94.0	92.6	94.0
CTG	82.1	83.6	82.1	82.7	83.0
Ionosphere	81.8	89.5	86.0	85.2	88.0
Krvskp	84.1	92.8	93.8	90.4	93.4
MADELON	59.2	63.3	53.8	56.0	62.2
PageBlock	89.4	93.0	91.2	91.4	89.1
Penbased	85.8	80.1	81.5	83.7	80.1
Texture	77.4	78.9	80.4	80.2	80.9
Waveform	78.8	70.7	74.3	79.1	77.0
Avg.	82.3	84.1	83.1	83.5	**84.3**

TABLE V. CLASSIFICATION ACCURACY OF C4.5 WITH AND WITHOUT FEATURE SELECTION (%)

Data set	**Orig.**	**FSDC (DP)**	**FSDC (HC)**	**FSDC (SOM)**	**FSDC (kMeans)**
Breast	93.7	94.4	92.6	93.3	93.5
Congress	96.3	94.9	95.4	95.2	95.4
CTG	93.2	90.7	90.0	92.0	90.4
Ionosphere	90.0	90.9	89.4	89.7	90.0
MADELON	75.5	70.3	76.5	74.5	83.3
Krvskp	99.3	93.6	93.8	98.0	94.2
PageBlock	97.0	96.5	96.3	96.5	96.1
Penbased	96.3	93.5	95.2	94.7	94.4
Texture	92.8	90.0	91.3	90.1	90.4
Waveform	74.4	72.3	74.6	77.8	75.9
Avg.	**90.8**	88.7	89.5	90.2	90.4

Experimental results confirm our assumption that we can get similar or better results with fewer features selected by the proposed methods. On nine of the ten benchmark data sets, naïve Bayes performs better on a small feature subset selected by the proposed methods than on the whole feature set (Table III). On seven of the ten benchmark data sets, kNN performs better with feature selection than without feature selection (Table IV). C4.5 performs similar or better on the small feature subsets selected by the proposed methods than on the whole feature sets (Table V).

The number of features selected by a method for different data sets is different (Table II), which means that the optimal number of features is different for different data sets and it is difficult to determine it in a prior. In addition, the number of features selected by the proposed methods are much fewer than that in the original set (Table II), which means that the intrinsic dimensionality of these data sets are small.

C. *Experiment 2:*

The goal of this experiment is to compare the proposed feature selection algorithms with several state-of-the-art feature selection methods.

Experimental design:

Few of feature selection methods can automatically determine the number of features to select as the proposed methods. We use the three methods we found in the literature as the comparison methods, including IAMB (Incremental Association Markov Blanket) proposed by Tsamardinos et al.[18], FCBF (Fast Correlation-Based Filter) proposed by Yu and Liu [19], and CMI (Conditional Mutual Information) proposed by Brown et al.[20]. Using the ten benchmark data sets, we evaluate the performance of feature selection methods by the classification accuracy of different classifiers, including kNN, naïve Bayes, and C4.5.

Experimental results:

Table VI shows the average number of feature selected by different methods. Fig.2 shows the average accuracies of different classifiers with various feature selection methods.

Among the different combinations of classifier and feature selection methods, FSDC(SOM) with kNN performs the best. FSDC(kMeans) performs well when combined with all the three classifiers compared with the other feature selection methods. FCBF performs well when combined with naïve Bayes and CMI performs well when combined with C4.5. FSDC(DP) and FCBC tends to select fewer features, while CMI tends to retain more features which may contains some redundant features.

TABLE VI. AVERAGE NUMBER OF FEATURE SELECTED BY DIFFERENT METHODS

Org.	**FCBF**	**IAMB**	**CMI**	**FSDC (DP)**	**FSDC (HC)**	**FSDC (kMeans)**	**FSDC (SOM)**
74.3	4.8	6.4	11.4	**4.6**	6.6	5.7	6.4

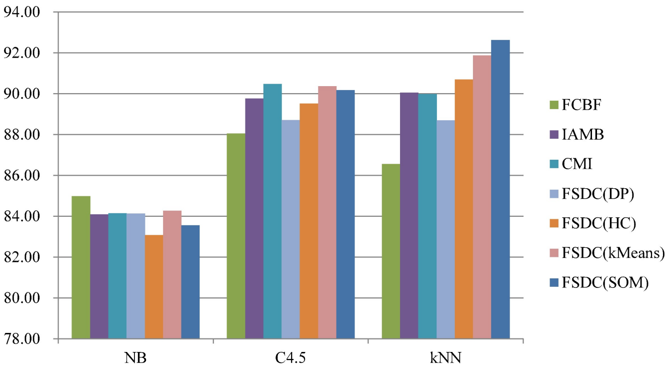

Fig. 2. Average accuracies of different classifiers with various feature selection methods

V. CONCLUSIONS

In this paper, we propose an information-theoretic feature selection framework based on supervised data clustering. It directly estimates the high-dimensional mutual information through supervised data clustering, which removes the assumptions that there only exists lower-order dependencies between features and the target variable. Experimental results show that: (1) the proposed method could be used to avoid the curse of dimensionality, especially for naïve Bayes and kNN, i.e. getting better performance with fewer features; (2) the proposed methods performs similar or better compared with state-of-the-art feature selection methods, especially when combined with kNN classifier.

REFERENCES

[1] I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh, eds., Feature Extraction: Foundations and Applications. Studies in Fuzziness and Soft Computing. Springer, 2006.

[2] I. Guyon and A. Elissee, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, pp.1157–1182, 2003.

[3] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Norwell, MA, USA: Kluwer Academic Publishers , 1998.

[4] J. R. Vergara and P.A. Estevez, “A review of feature selection methods based on mutual information,” Neural Computing and Applications, vol. 24, no. 1, pp. 175–186, 2014.

[5] H. Liu, Z. Wu, X. Zhang, and D. F. Hsu. "An information-theoretic feature selection method based on estimation of Markov blanket." In 14th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC’15), pp. 327-332, 2015.

[6] G. Brown, “A new perspective for information theoretic feature selection,” in AISTATS’09, pp.49–56, 2009.

[7] H. Liu, Z. Wu, and X. Zhang. "Feature Selection Based on Data Clustering." In Intelligent Computing Theories and Methodologies, Volume 9225 of Lecture Notes in Computer Science, pp. 227-236, 2015.

[8] D. Pál, B. Póczos, and C. Szepesvári. "Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs." In Advances in Neural Information Processing Systems (NIPS’10), pp. 1849-1857. 2010.

[9] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” IEEE Transactions on Neural Networks, vol.5(4), pp. 537–550, 1994.

[10] 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.

[11] K. S. Balagani, and V.V. Phoha, “On the feature selection criterion based on an approximation of multidimensional mutual information,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, no.7, pp.1342–1343, 2010.

[12] Q. Song, J. Ni, and G. Wang. "A fast clustering-based feature subset selection algorithm for high-dimensional data." IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp. 1-14, 2013.

[13] Hellman, M., Raviv, J.: Probability of error, equivocation, and the chernoff bound. IEEE Transactions on Information Theory 16(4):368–372 (1970)

[14] A. Rodriguez, and A. Laio, “Clustering by fast search and find of density peaks,” Science, Vol. 344, no. 6191, pp.1492–1496, 2014.

[15] S.P. Lloyd, “Least squares quantization in PCM”. Information Theory, IEEE Transactions on, 28(2): 129-137, 1982.

[16] T. Kohonen, "The self-organizing map," Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480, 1990.

[17] K. Bache and M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml.

[18] I. Tsamardinos, C. F. Aliferis, and A. Statnikov. “Algorithms for large scale markov blanket discovery,” In 16th International FLAIRS Conference, vol. 103, 2003.

[19] L. Yu, and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” Journal of Machine Learning Research, Vol. 5, pp.1205-1224, 2004.

[20] G. Brown, A. Pocock, M.J. Zhao, and M. Lujan, “Conditional likelihood maximisation: A unifying framework for information theoretic feature selection,” Journal of Machine Learning Research, Vol.13, pp. 27–66 , 2012.