Simultaneous Feature Selection and Feature Extraction for Pattern Classification

a dissertation submitted in partial fulfilment of the requirements for the M.Tech. (Computer Science) degree of Indian Statistical Institute

By

Sreevani
Roll No. Cs0723

under the supervision of

Prof C A Murthy
Machine Intelligence Unit

I N D I A N S T A T I S T I C A L I N S T I T U T E

203, Barrackpore Trunk Road

Kolkata - 700108

Indian Statistical Institute

CERTIFICATE

This is to certify that the thesis entitled *'Simultaneous Feature Selection and Feature Extraction for Pattern Classification'* is submitted in the partial fulfilment of the degree of M. Tech. in Computer Science at Indian Statistical Institute, Kolkata.

The work carried out by Sreevani under my supervision and guidance, is fully adequate, in scope and quality as a dissertation for the required degree.
It is further certified that no part of this thesis has been submitted to any other university or institute for the award of any degree or diploma.

Prof. C A Murthy
(Supervisor)

Countersigned
(External Examiner)
Date: of July 2009

Contents

Acknowledgement

I take this opportunity to thank **Prof. C A Murthy**, Machine Intelligence-Unit, ISI-Kolkata for his valuable guidance, inspiration. His pleasant and encouraging words have always kept my spirits up. I am greatful to him for providing me to work under his supervision. Also, I am very thankful to him for giving me the idea behind the algorithm developed in this thesis work.

Finally I would like to thank all of my colleagues, class mates, friends, and my family members for their support and motivation to complete this project.

Sreevani

M. Tech(CS)

Declaration

No part of this thesis has been previously submitted by the author for a degree at any other university, and all the results contained within, unless otherwise stated are claimed as original. The results obtained in this thesis are all due to fresh work.

Sreevani

Date

Abstract

Feature subset selection and extraction algorithms are actively and extensively studied in machine learning literature to reduce the high dimensionality of feature space, since high dimensional data sets are generally not efficiently and effectively handled by machine learning and pattern recognition algorithms. In this thesis, a novel approach to combining feature selection and feature extraction algorithms. We present an algorithm which incorporates both. The performance of the algorithm is established over real-life data sets of different sizes and dimensions.

Chapter 1. Introduction

With advanced computer technologies and their omnipresent usage, data accumulates in a speed unmatchable by the human's capacity to process it. To meet this growing challenge, the research community of knowledge discovery from databases emerged. The key issue studied by this community is to make advantageous use of large stores of data. In order to make raw data useful, it is necessary to represent, process, and extract knowledge for various applications. For searching meaningful patterns in raw data sets, Data Mining algorithms are used.

Data with many attributes generally presents processing challenges for data mining algorithms[15]. Model attributes are the dimensions of the processing space used by the algorithm. The higher the dimensionality of the processing space, the higher the computation cost involved in algorithmic processing. Dimensionality constitutes a serious obstacle to the efficiency of most Data Mining algorithms. Irrelevant attributes simply add noise to the data and affect model accuracy. Noise increases the size of the model and the time and system resources needed for model building and scoring. Moreover, data sets with many attributes may contain groups of attributes that are correlated. These attributes may actually be measuring the same underlying feature. Their presence together in the data can skew the logic of the algorithm and affect the accuracy of the model.

To minimize the effects of noise, correlation, and high dimensionality, some form of dimension reduction is sometimes a desirable preprocessing step for data mining. Feature selection and extraction are two approaches to dimension reduction. Due to increasing demands for dimensionality reduction, research on feature selection/Extraction has deeply

and widely expanded into many fields, including computational statistics, pattern recognition, machine learning, data mining, and knowledge discovery.

Feature selection/Extraction is an important problem, not only for the insight gained from determining relevant modeling variables, but also for the improved understandability, scalability, and, possibly, accuracy of the resulting models.

Feature Selection — Selecting the most relevant attributes .

Feature Extraction — Combining attributes into a new reduced set of features.

1.2 Brief Overview of Related Work

Feature selection and feature extraction algorithms reduce the dimensionality of feature space while preserving enough information for the underlying learning problem. There are a plenty of feature subset selection/extraction algorithms proposed in the machine learning literature. In this section, we will introduce some of the Feature selection/Extraction algorithms proposed in the literature.

1.2.1. Feature Selection:

Feature selection is one effective means to identify relevant features for dimension reduction. Various studies show that features can be removed without performance deterioration. The training data can be either labeled or unlabeled, leading to the development of supervised and unsupervised feature selection algorithms. To date, researchers have studied the two types of feature selection algorithms largely separately. Supervised feature selection determines feature relevance by evaluating feature's correlation with the class, and without labels, unsupervised feature selection exploits data variance and separability to evaluate feature relevance. In general, feature selection is a search problem according to some evaluation criterion.

Feature selection is a process that chooses a subset of M features from the original set of N features ($M \leq N$), so that the feature space is optimally reduced according to a certain criterion[3,5].

Feature subset generation:

One intuitive way is to generate subsets of features sequentially. If we start with an empty subset and gradually add one feature at a time, we adopt a scheme called sequential forward selection; if we start with a full set and remove one feature at a time, we have a scheme called sequential backward selection. We can also randomly generate a subset so that each possible subset (in total, 2^N, where N is the number of features) has an approximately equal chance to be generated. One extreme way is to exhaustively enumerate 2^N, possible subsets.

Feature evaluation:

An optimal subset is always relative to a certain evaluation criterion (i.e. an optimal subset chosen using one evaluation criterion may not be the same as that using another evaluation criterion). Evaluation criteria can be broadly categorized into two groups based on their dependence on the learning algorithm applied on the selected feature subset. Typically, an independent criterion (i.e. filter) tries to evaluate the goodness of a feature or feature subset without the involvement of a learning algorithm in this process. Some of the independent criteria are distance measure, information measure, dependency measure. A dependent criterion (i.e. wrapper) tries to evaluate the goodness of a feature or feature subset by evaluating the performance of the learning algorithm applied on the selected subset. In other words, it is the same measure on the performance of the applied learning algorithm. For supervised learning, the primary goal of classification is to maximize predictive accuracy, therefore, predictive accuracy is generally accepted and widely used as the primary measure by researchers and practitioners. While for unsupervised learning, there exist a number of heuristic criteria for estimating the quality of clustering results, such as cluster compactness, scatter separability, and maximum likelihood.

Algorithms :

Many feature selection algorithms exist in the literature.

Exhaustive/complete approaches :

Exhaustively evaluates subsets starting from subsets with one feature (i.e., sequential forward search); Branch-and-Bound [2] evaluates estimated accuracy, and ABB [22] checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.

Heuristic approaches :

SFS (sequential forward search) and SBS (sequential backward search) [2] can apply any of five measures. DTM [4] is the simplest version of a wrapper model - just learn a classifier once and use whatever features found in the classifier.

Nondeterministic approaches :

LVF [18] and LVW [19] randomly generate feature subsets but test them differently: LVF applies an inconsistency measure, LVW uses accuracy estimated by a classifier. Genetic Algorithms and Simulated Annealing are also used in feature selection [30, 13]. The former may produce multiple subsets, the latter produces a single subset.

Instance based approaches :

Relief [15, 16] is a typical example for this category. There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.

1.2.2. Feature Extraction

Feature extraction is a process that extracts a set of new features from the original features through some functional mapping [17]. Assuming there are n features (or attributes) $A_1, A_2, \ldots, A_n$, after feature extraction, we have another set of new features $B_1, B_2, ..., B_n$,

$B_i = F_i$ ($A_1; A_2$; :::; A_n), and F_i is a mapping function. Intensive search is generally required in finding good transformations. The goal of feature extraction is to search for a minimum set of new features via some transformation according to some performance measure. The major research issues can therefore be summarized as follows.

Performance Measure :

It investigates what is the most suitable in evaluating extracted features. For a task of classification, the data has class labels and predictive accuracy might be used to determine what is a set of extracted features. When it is of clustering, the data does not have class labels and one has to resort to other measures such as inter-cluster/intracluster similarity, variance among data, etc.

Transformation :

It studies ways of mapping original attributes to new features. Different mappings can be employed to extract features. In general, the mappings can be categorized into linear or nonlinear transformations. One could categorize transformations along two dimensions: linear and labeled, linear and non-labeled, nonlinear and labeled, nonlinear and non-labeled. Many data mining techniques can be used in transformation such as EM, k-Means, k-Medoids, Multi-layer Perceptrons, etc [12].

Number of new features:

It surveys methods that determine the minimum number of new features. With our objective to create a minimum set of new features, the real question is how many new features can ensure that ``the true nature'' of the data remains after transformation.

One can take advantage of data characteristics as a critical constraint in selecting performance measure, number of new features, and transformation. In addition to with/without class labels, data attributes can be of various types: continuous, nominal, binary, mixed. Feature extraction can find its many usages: dimensionality reduction for further processing [23], visualization [8], compound features used to booster some data mining algorithms [20].

Algorithms:

The functional mapping can be realized in several ways. We present here two exemplar algorithms to illustrate how they treat different aspects of feature extraction.

A feed forward neural networks approach :

A single hidden layer multilayer perceptron can be used to to extract new features [29]. The basic idea is to use the hidden units as newly extracted features. The predictive accuracy is estimated and used as the performance measure. This entails that data should be labeled with classes. The transformation from input

units to hidden units is non-linear. Two algorithms are designed to construct a network with the minimum number of hidden units and the minimum of connections between the input and hidden layers: the network construction algorithm parsimoniously adds one more hidden unit to improve predictive accuracy; and the network pruning algorithm generously removes redundant connections between the input and hidden layers if predictive accuracy does not deteriorate.

Principal Component Analysis:

PCA is a classic technique in which the original n attributes are replaced by another set of m new features that are formed from linear combinations of the original attributes. The basic idea is straightforward: to form an m-dimensional projection ($1 \leq m \leq n-1$) by those linear combinations that maximize the sample variance subject to being uncorrelated with all these already selected linear combinations. Performance measure is sample variance; the number of new features, m, is determined by the m principal components that capture the amount of variance subject to a pre-determined threshold; and the transformation is linear combination. PCA does not require that data be labeled with classes. The search for m principal components can be rephrased to finding m eigenvectors associated with the m largest eigenvalues of the covariance matrix of a data set [12].

One drawback of feature extraction algorithms when compared to feature selection algorithms is that the new features created by feature extraction algorithms may not carry any physical meaning of original features. Thus when design new feature extraction algorithms, it makes sense to maintain the meaning of original features in some ways.

1.3 Objective of the work

Feature Selection/Extraction have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. To reduce high dimensionality of the data, there are methods which involve only either feature selection or feature extraction. These methods provide reduced feature subset with only either original features or transformed features. Sometimes it may be useful to have both original and transformed features in the reduced feature set. Can we have one such method which provides reduced feature subset with original as well as transformed features? Can we have a method which incorporates both feature selection and feature extraction to get the best of two? The goal of this project is to develop a method which incorporates both feature selection and extraction to reduce high dimensionality. This work is an attempt to unify both feature selection and extraction methods.

1.4 Organisation of the report

The organization of the report is as follows:

In Chapter 2, the proposed method is discussed. Algorithm for the proposed method also its computational complexity is discussed. In chapter 3, different data sets, methods, feature evaluation indices used for comparison are discussed. In Chapter 4, experimental results along with comparisons are provided. Last Chapter, Chapter 5 concludes the discussions of *'Simulataneous feature selection and feature extraction for pattern classification'.*

Chapter 2 : Proposed Method

2.1. Motivation:

The chasm between Feature Selection and Feature Extraction seems difficult to close as one works with Original features and the other works with Transformed features. However, if we change the perspective and put less focus on type of features(either original/transformed), both Feature Selection and Feature Extraction can be viewed as an effort to get features that are consistent with the target concept. In Feature selection the target concept is related to original features, while in feature extraction the target concept is usually related to transformed features. Essentially, in both cases, the target concept is related to reducing the dimensionality so that the classification success rate is high. The challenge now is how to develop a unified method which combines Feature Selection and Feature Extraction for pattern classification.

2.2. Outline of the proposed algorithm

The task of Dimensionality Reduction for pattern classification involves, at each step, two stages. First stage involves feature selction and second stage involves feature extraction.

In each stage, we consider all possible pairs of features.

In the First stage, from each pair of random variables , the feature carrying little or no additional information beyond that carried by the other feature ,is redundant and it is eliminated. For this purpose, we have used the dependency measure, correlation coefficient (ρ). If the value of the ρ between a pair is greater than a predefined threshold, then the feature with minimum variance is discarded and the one with maximum feature is retained to constitute the reduced feature set.

In the Second stage, from each pair of random variables, one feature is extracted if the error introduced is considerable i.e. first principal component of the pair is taken to form reduced feature set and the pair is discarded. Since λ_2, smallest eigen value of the covariance matrix $\sum$ of the above pair, provides the variance along the second principal component, error can be taken as λ_2. So, λ_2 is the amount of error introduced while projecting the data to reduced dimension (here from two to one) in a best possible way. Since $'\lambda_2'$ is arbitrary, it can be normalized by sum of the eigen values of $\sum$. So, we extract a feature from each pair if the error e = $\frac{\lambda_2}{\lambda_1+\lambda_2}$ introduced is considerable i.e. if the value of 'e' is below a predefined threshold then we extract a feature from these two features and the pair of features is discarded.

The process is repeated for the remaining features until all of them are selected or discarded.

2.3. Algorithm

Let the original number of features be D, and the original feature set be O = $\{F_i , i = 1,2,\ldots,D\}$. Let T_1, T_2 be two predefined threshold values.

Step 1 : Intialize the reduced feature subset R to the original feature set O.

Step 2 : for each pair of features F_i, F_j in O,

If $\rho(F_i, F_j) > T_1$ then

Retain the feature with maximum variance say F_i and discard the feature with lower variance say F_j.

Step 3 : for each pair of features F_i, F_j in O,

If $e(F_i, F_j) < T_2$ then

Add the linear combination of the features F_i, F_j to the reduced feature set and discard both F_i and F_j.

Step 4 : If (|R| == |O|) then

Go To Step 5.

Else

Rename ‘R’ as ‘O’ and **Go to Step 2.**

Step 5 : Return feature set R as the reduced feature set.

2.4. Computational Complexity:

Since in each stage every pair of features is considered, with respect to dimension(*D*), the proposed algorithm has complexity $\mathcal{O}(D^2)$, where *D* is the number of original features.

If the data set contains ℓ number of samples, evaluation of correlation coefficient(ρ) and error(e), defines as above, for a feature pair is of complexity $\mathcal{O}(\ell)$.

Thus, the proposed method has overall complexity $\mathcal{O}(D^2\ell)$.

Chapter 3 :

Data Sets, Methods, Feature Evaluation Indices used for Comparison

In this Chapter, we present different data sets, Methods, Evaluation indices used for comparison are discussed. First, the characteristics of the data sets used are discussed. Second, different methods used for comparison are described. Third, Feature Evaluation Indices to compare the performance of the proposed algorthim with the other methods are described.

3.1. Data Sets

Three categories of real-life public domain data sets are used: low-dimensional ($D \leq 10$), medium-dimensional($10 < D \leq 100$), and high-dimensional ($D > 100$), containing both large and relatively smaller number of points. They are available from the UCI Machine Learning Repository.

1. ***Isolet*** **.** The data consists of several spectral coefficients of utterances of English alphabets by 150 subjects. There are 617 features all real in the range [0,1], 7,797 instances, and 26 classes.

2. ***Multiple Features*** **.** This data set consists of features of Hand written numerals (0-9) extracted from a collection of Dutch utility maps. There are total 2,000 patterns, 649 features, and 10 classes.

3. ***Spambase* .** The task is to classify an email into spam or nonspam category. There are 4,601 instances, 57 continuous valued attributes denoting word frequencies, and 2 classes.

4.***Ionosphere.*** The data represents autocorrelation functions of radar measurements. The task is to classify them into two classes denoting passage or obstruction in ionosphere. There are 351 instances and 34 attributes, all continuous.

5. ***Wisconsin Cancer.*** The popular Wisconsin breast cancer data set contains nine features, 684 instances, and two classes.

6. ***Iris.*** The data set contains 150 instances, four features, and three classes of Iris flowers.

7. ***Waveform***. This consists of 5,000 instances having 40 attributes each. The attributes are continuous valued, and some of them are noise. The task is to classify an instance into one three categories of waves.

3.2. Methods

1. ***Branch and Bound Algorithm (BB)* .** A search method in which all possible subsets are implicitly inspected without exhaustive search. If the feature selection criterion is monotonic BB returns the optimal subset.

2. ***Sequential Forward Search (SFS).*** A suboptimal search procedure where one feature at a time is added to the current feature set. At each stage, the Feature to be included in the feature set is selected from among the remaining available features so that the new enlarged feature set yields a maximum value of the criterion function used.

3. ***Unsupervised feature selection Using feature similarity.*** A method in which the original features are partitioned into a number of homogeneous subsets based on the KNN principle

using the similarity measure '*maximal information compression index*'. Among them the feature having the most compact subset is selected, and its k nearest neighbors are discarded.

3.3. Feature Evaluation Indices:

Some indices that have been considered for evaluating the effectiveness of the selected feature subsets are given below. The first two indices, namely class seperability, K-NN misclassification rate, do need class information of the samples while the remaining one namely, Entropy, do not.

Notation:

Let ℓ be the number of sample points in the data set, *c* be the number of classes present in the data set, *D* be the number of features in the original feature set *O*, *d* be the number of features in the reduced feature set *R*, Ω_o be the original feature space with dimension D, and Ω_R be the transformed feature space with dimension d.

1. KNN misclassification Rate. It is used for evaluating the effectiveness of the reduced set for classification. Cross-validation is performed in the following manner.

We randomly select 50 percent of the data as training set and classify the remaining 50 percent points. Ten such independent runs are performed and the average classification accuracy on test set is used. The value of K, for the KNN rule, is equal to 1.

2.Class Separability. Class Separability S of a data set is defined as S = *trace*($S_b^{-1}S_w$). S_w is the within class scatter matrix and S_b is the between class scatter matrix, defined as:

$$S_w = \sum_{j=1}^{c} \Pi_j \, E\{(X-\mu_j)(X-\mu_j)^T \mid \omega_j\} = \sum_{j=1}^{c} \Pi_j \sum_j$$

$$S_b = \sum_{j=1}^{c} (\mu_j - M_o)(\mu_j - M_o)^T$$

$$M_o = E\{X\} = \sum_{j=1}^{c} \Pi_j \mu_j$$

Where Π_j is the a priori probability that a pattern belongs to a class

ω_j, X is the feature vector, μ_j is the sample mean vector of class ω_j, M_o is the sample mean vector for the entire data points, $\sum_j$ is the sample covariance matrix of class ω_j, and $E\{.\}$ is the expectation operator. A lower value of the separability criteria *S* ensures that the classes are well separated by their scatter means.

3. Entropy. Let the distance between two data points p, q be

$$\mathcal{D}_{pq} = \left[\sum_{j=1}^{M} \left(\frac{x_{p,j} - x_{q,j}}{max_j - min_j}\right)^2\right]^{1/2}$$

Where $x_{p,j}$ denotes feature value for p along j th direction, and max_j, min_j are the maximum and minimum values computed over all the samples along the j th direction, M is the number of features. Similarity between p, q is given by $sim(p,q) = e^{-\alpha \mathcal{D}_{pq}}$, where α is a positive constant. A possible value of α is $\frac{-\ln 0.5}{\bar{\mathcal{D}}}$. $\bar{\mathcal{D}}$ is the average distance between data points computed over the entire data set. Entropy is defined as:

$$E = \sum_{p=1}^{t} \sum_{q=1}^{t} (sim(p,q) \times \log sim(p,q) + \big(1 - sim(p,q)\big) \times \log(1 - sim(p,q))).$$

If the data is uniformly distributed in the feature space, entropy is maximum. When the data has well formed clusters uncertainty is low and so is entropy.

Chapter 4 :

Experimental Results and Comparisons

Performance of the proposed algorithm in terms of the feature evaluation indices, discussed in the previous chapter, is compared with three other feature selection schemes. Also, Effect of varying parameters T_1, T_2 , used as thresholds, is also given.

Comparative results are shown for large, medium, low dimensional data sets in terms of feature evaluation indices which are described in the last chapter. Table 1, Table 2, Table 3 provides the comparative result for high, medium, low dimensional data sets respectively. For the misclassification rates using KNN, mean and standard deviation computed for ten independent runs are provided.

4.1. Classification and Clustering Performance

Classification Performance of the proposed algorithm in terms of two indices namely KNN misclassification rate, Class Separability is slightly higher than the other three mentioned algorithms.

Table 1: Comparative results for high dimensional data sets

Data Set	Method	Evaluation Criteria			
		KNNM		S	E
		Mean	SD		
Isolet I = 7797, C = 26 D = 617, d = 383	**SFS**	0.09	0.02e-005	1.5	0.58
	UFS using feature similarity	0.08	0.62e-004	1.45	0.55
	Proposed	0.07	1.82e-005	1.40	0.58
Multiple Features I = 2000, C=10 D=64 9, d= 328	**SFS**	0.06	1.40e-005	0.65	0.68
	UFS using feature similarity	0.05	3.40e-005	0.65	0.70
	Proposed	0.03	9.51e-006	0.55	0.75

SFS : Sequential Forward Search, UFS : Unsupervised Feature Selection, KNNM : k-Nearest Neighborhood Misclassification Rate , S : Separability, E : Entropy

Table2 : Comparative results for medium dimensional data sets

Data Set	Method	Evaluation Criteria			
		KNNM		S	E
		Mean	SD		
Spambase I =4 601, C=2 D= 57, d = 34	**BB**	0.22	4.86e-005	2.83	0.60
	SFS	0.23	3.06e-005	2.95	0.65
	UFS using feature similarity	0.22	4.86e-005	2.83	0.60
	Proposed	0.11	3.54e-005	1.42	0.64
Waveform I = 5000, C=3 D=40, d = 24	**BB**	0.22	1.82e-005	0.38	0.76
	SFS	0.26	2.14e-004	0.48	0.79
	UFS using feature similarity	0.25	2.12e-003	0.40	0.77
	Proposed	0.25	1.32e-003	0.38	0.80
Ionosphere I = 351, C=2 D=34, d = 17	**BB**	0.18	2.86e-005	0.25	0.76
	SFS	0.21	1.06e-003	0.30	0.76
	UFS using feature similarity	0.16	5.95e-004	0.30	0.75
	Proposed	0.15	5.48e-004	0.25	0.77

BB: Branch and Bound, SFS : Sequential Forward Search, UFS : Unsupervised Feature selection, KNNM : k-Nearest Neighborhood Misclassification rate , S : Separability, E : Entropy.

Table 3 : Comparative results for low dimensional data sets

Data Set	Method	Evaluation Criteria			
		KNNM		S	E
		Mean	SD		
Cancer I = 683, C=2 D= 9, d = 4	**BB**	0.06	2.17e-004	1.84	0.46
	SFS	0.07	1.17e-004	2.68	0.48
	UFS using feature similarity	0.05	2.17e-005	1.70	0.43
	Proposed	0.04	9.3476	1.70	0.50
Iris I = 150, C=3 D= 4, d = 2	**BB**	0.07	3.18e-004	20.0	0.55
	SFS	0.08	1.19e-003	25.0	0.57
	UFS using feature similarity	0.07	3.18e-004	20.0	0.55
	Proposed	0.07	5.45e-004	20.0	0.60

BB: Branch and Bound, SFS : Sequential Forward Search, UFS : Unsupervised Feature selection, KNNM : k-Nearest Neighborhood Misclassification rate , S : Separability, E : Entropy

Choice of T_1, T_2 :

In this algorithm T_1, T_2 controls the size of the reduced set. We can as many features in the reduced data set as want, by setting appropriate values to thresholds. Since T_1, T_2 determine the error thresholds, data representation is controlled by their choice.

Chapter 5 : Conclusion & Scope

In this thesis, an algorithm for dimensionality reduction is described. Feature selection and feature extraction are mainly two approaches to mining large data sets, both in dimension and size. Here, we have proposed an algorithm which combines both feature selection and feature extraction for dimensionality reduction to get the best of two. This is a way of unifying both feature selection and feature extraction. This algorithm uses correlation coefficient as the dependency measure between a pair of features.

As mentioned above, the proposed method is a way of performing both feature selection and feature extraction simultaneously. One can develop a different way of combining both feature selection/extraction.

Refernces

[1] Pabitra Mitra, C A Murthy and Sankar K Pal, “Unsupervised Feature Selelction Using Feature Similarity”, IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 24, No 3, March 2002.

[2] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs: Prentice Hall, 1982.

[3] A.L. Blum and P. Langley, ``Selection of Relevant Features and Examples in Machine Learning", Artificial Intelligence, 97, pp. 245--271, 1997.

[4] C. Cardie, ``Using Decision Trees to Improve Case-Based Learning", Proc of the Tenth International Conference on Machine Learning, pp. 25-- 32, 1993.

[5] M. Dash and H. Liu, ``Feature Selection Methods for Classifications", Intelligent Data Analysis: An International Journal, 1, 3, 1997.http://www-east.elsevier.com/ida/free.-htm.

[6] U. Fayyad and R. Uthurusamy, “Data Mining and Knowledge Discovery in Databases”, Comm. ACM, vol. 39, no. 11, pp. 24-27, Nov. 1996.

[7] M. Kudo and J. Sklansky, “Comparison of Algorithms that Selects Features for Pattern Classifiers”, Pattern Recognition, vol. 33, pp. 25-41, 2000.

[8] Oded Maimon and Lior Rokach, “The Data Mining and Knowledge Discovery Handbook”, Tel-Aviv University, Israel.

[7] J. G. Dy and C. E. Brodley, ``Feature Subset Se- lection and Order Identification for Unsupervised Learning”, Proc. of the Seventeenth International Conference on Machine Learning, pp. 247--254, 2000.

[8] M. Dash and H. Liu, ``Unsupervised Feature Selection Methods", *proc. Pacific*

Asia Conf. Knowledge Discovery and Data Mining, pp. 110-121,2000.

[9] P. Pudil, J. NovovicovaÂ, and J. Kittler, "Floating Search Methods in Feature Selection", Pattern Recognition Letters, vol. 15, pp. 1119-1125, 1994.

[10] DW. Aha and R.L. Bankert, "A Comparative Evaluation of Sequential Feature Selection Algorithms", Artificial Intelligence and Statistics V, D. Fisher and J.-H. Lenz, eds., New York: Springer Verlag, 1996.

[11] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufman, 2001.

[12] K. Kira and L. Rendell, ``The feature selection problem: Traditional methods and a new algorithm", Proc. of the Tenth National Conference on Artificial Intelligence, pp. 129--134, 1992.

[13] D. Koller and M. Sahami, "Towards Optimal Feature Selection", 13th Int'l. Conf. Machine Learning, pp. 284-292, 1996.

[14] H. Liu and R. Setiono, "Some Issues in Scalable Feature Selection",Expert Systems with Applications, vol. 15, pp. 333-339, 1998.

[15] H. Liu and H. Motoda, editors, "Feature Extraction, Construction and Selection: A Data Mining Perspective", Boston: Kluwer Academic Publishers, 1998. 2nd Printing, 2001.

[16] Z. Zheng, ``A Comparison of Constructing Different Types of New Features for Decision Tree Learning", chapter 15, pp. 239 -- 255. In [23], 1998. 2nd Printing, 2001.

[17] N. Wyse, R. Dubes, and A.K. Jain, ``A critical evaluation of intrinsic dimensionality algorithms", In E.S. Gelsema and L.N. Kanal, editors, Pattern Recognition in Practice, pp. 415-- 425. Morgan Kaufmann Publishers, Inc., 1980.