Dimension Reduction Methods in High Dimensional Data Mining

Pramod Kumar Singh

Associate Professor

ABV-Indian Institute of Information Technology and Management Gwalior

Gwalior – 474015, Madhya Pradesh, India

Outline

- Introduction
- Feature Selection
- Feature Extraction
- Hybrid Methods
- Feature Selection Using Multi-objective Optimization Methods
- Summary
- References

Introduction: Classification and Clustering

- This presentation is focused with respect to two data mining tasks clustering and classification.
- Classification
  - Classification classifies the labeled data set based on supervised learning.
- Clustering
  - Clustering groups the unlabeled data set based on unsupervised learning.

Introduction – Dimension / Feature

Dataset

A dataset is a collection of homogeneous objects.

Object

An instance in the dataset is known as an object.

Dimensions / Features

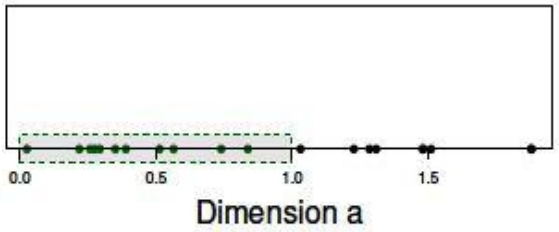
Properties, which define an object, are known as dimensions / features. It separates one object from the other.

Dataset

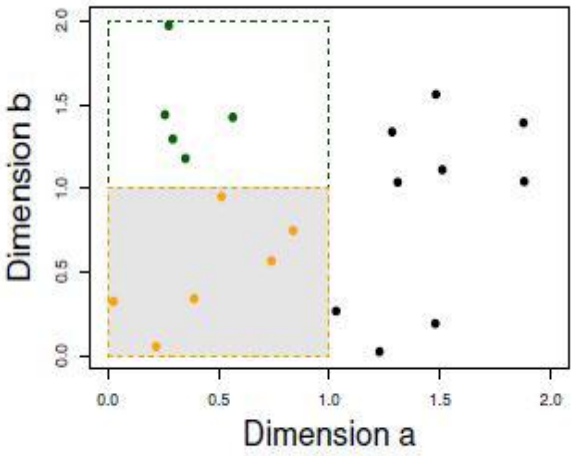
Feature 1	Feature 2	Feature 3	…	Feature n
1.1	Red	0		Engineer
2.5	Green	1		Manager
3.7	Blue	1		G.M.

← object

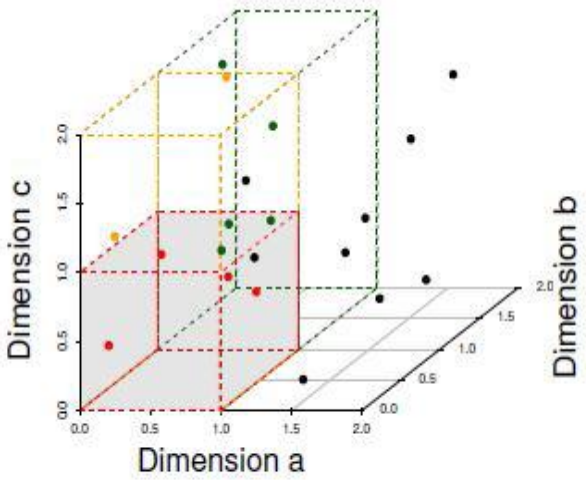
Introduction - Curse of Dimensionality



(a) 11 Objects in One Unit Bin



(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin

L. Parsons, E. Haque and H. Liu, "Subspace Clustering for High Dimensional Data: A Review", ACM SIGKDD Explorations Newsletter – Special Issue on Learning from Imbalanced Datasets, Vol. 6, Issue 1, pp. 90-105, June 2004.

Introduction- Curse of Dimensionality

High dimensions

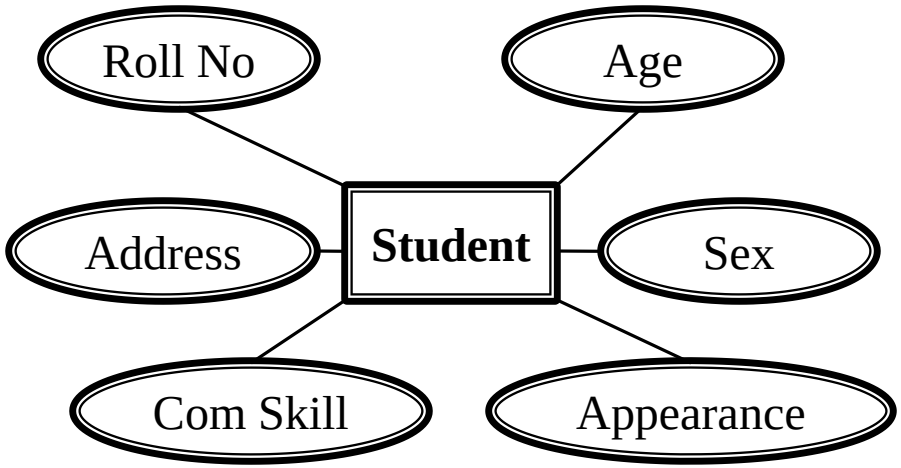
- Increases computational complexity of the underlying algorithm
- Reduces quality of the results and
- Sometimes misleads the algorithm

Because real-world large dataset consists of irrelevant, redundant, and noisy dimensions, we may consider dimension reduction.

Introduction - Dimension Reduction Methods

Data reduction methods

reduce dimensionality of the dataset to avoid the curse of dimensionality without substantial loss of information and without affecting the final output.



For personality

- Roll No is irrelevant
- Address is redundant

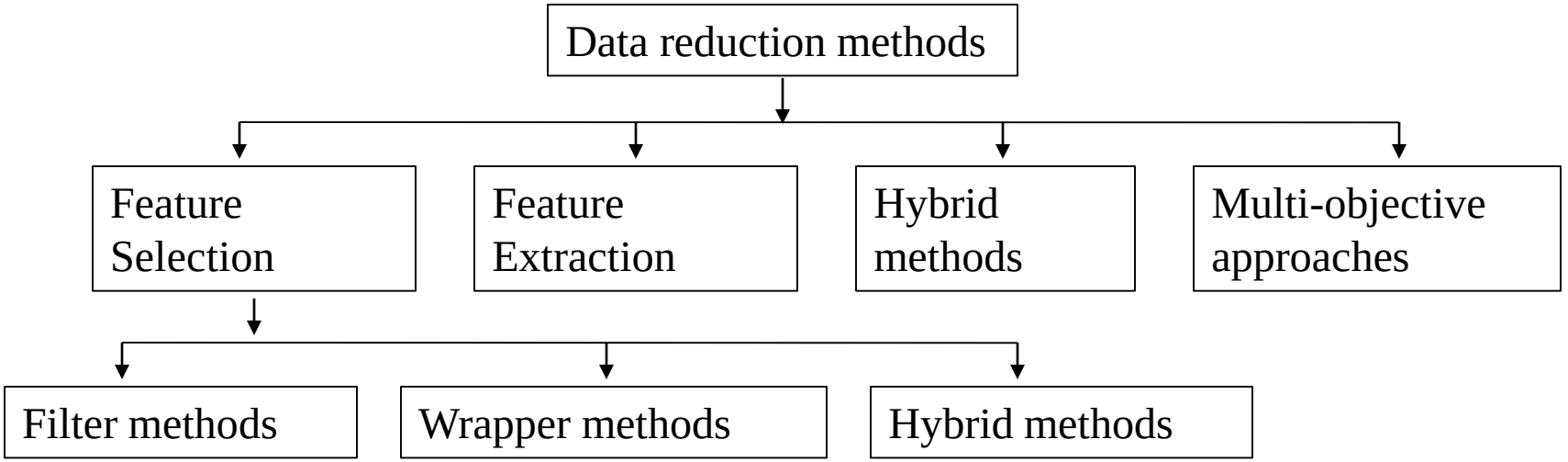
Introduction - Dimension Reduction Methods

Objectives of feature selection can be as follows

- Improving the performance of learning algorithm.
- Reducing the storage requirement for data set.
- Enhancing data understanding and helping to visualize it.
- Detecting noises and outliers in the data set.

Introduction: Data Reduction Methods

Following figure shows a tree structure of the data reduction methods.



Introduction: Feature Selection

- Feature selection is a process that selects a subset of original features by rejecting irrelevant and/or redundant features according to certain criteria.
- Relevancy of features is typically measured by discriminating ability of a feature to enhance predictive accuracy of classifier and cluster goodness for clustering algorithm.
- Generally, feature redundancy is defined by correlation; two features are redundant to each other if their values are correlated.

Introduction: Feature Selection (Contd.)

- Applications:
  - Feature selection is an innovative area of research in pattern recognition, machine learning, and data mining and is widely applied to many fields such as text categorization, image retrieval, customer relationship management , intrusion detection.

Introduction: Feature Extraction

- Feature extraction, an alternative approach of data reduction, is a special type of dimensionality reduction method to create a set of new reduced features based on some transformation function.

Introduction: Comparative Analysis of Dimensional Reduction Methods

Comparative analysis of Feature Selection and Feature Extraction methods

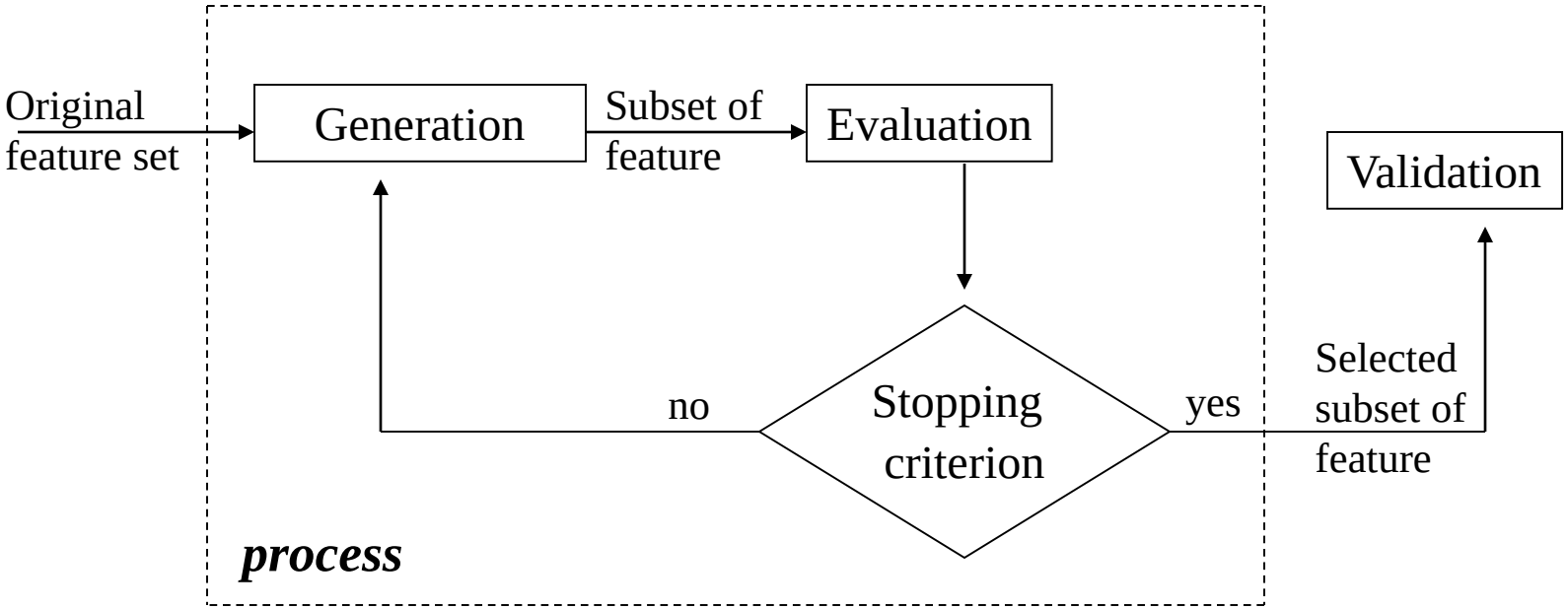
Name of approach	Key concept	Merits	Demerits	Example
Feature selection	Selects a subset of relevant features based on evaluation criteria	Robust against irrelevant feature	Criteria are hard to define and may generate redundant features	Information gain (IG)
Feature extraction	Reduces features by defined transformation function	Originality or relative distance between objects are preserved	Effectiveness in case of large number of irrelevant features is unsatisfactory	Principal component analysis (PCA)

Feature Selection

- Four key steps of feature selection
  - Subset generation
  - Subset evaluation
  - Stopping criteria
  - Result validation

Feature Selection: Common Steps

Description of common steps of feature selection



Generation = select feature subset candidate.
Evaluation = compute relevancy value of the subset.
Stopping criterion = determine whether subset is relevant.
Validation = verify subset validity.

Feature Selection: Subset Generation

- Feature selection is NP-hard even for moderate value of N as 2^N candidate subsets are possible for N original features in the data set.
- Subset generation is a search procedure that produces candidate feature subsets for evaluation based on a certain search strategy.
- Different available strategies can be categorized as complete, sequential, random, and heuristic search
- The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied.

Feature Selection: Subset Generation (Contd.)

- **Search strategies:**
  - Complete search guarantees to find the optimal result according to the evaluation criterion used through the backtracking.
  - Sequential search is a greedy hill-climbing approach. There are many variations of this approach, such as sequential forward selection (SFS), sequential backward elimination (SBE), bi-directional selection.
  - Random search starts with a randomly selected subset.
  - Heuristic search find the next subset based on some previous experience.

Feature Selection: Subset Generation (Contd.)

Comparison of search methods in subset generation

Method	Key concept	Merits	Demerits	Example
Complete	Search based on the concept of backtracking	Find optimal subset	Running time is exponential	Branch and Bound
Sequential	Removal of feature is sequential	Computational complexity is less	It may suffer from nesting effect	SFS,SBE
Random	Subset selection in random manner	skip to stuck into local optima	Absence of any type of prediction	Las vegus algorithm
Heuristic	Search based on previous experience	produce many candidate near optimal solutions	Computational complexity is high	Genetic algorithm

Feature Selection: Subset Evaluation

- Each newly generated subset needs to be evaluated by an evaluation criterion.
- Feature selection algorithms designed with different evaluation criteria are broadly categorized into three categories:
  - Filter model
  - Wrapper model
  - Hybrid model

Feature Selection: Subset Evaluation (Contd.)

Comparison of three key methods for feature selection

Criteria	Filter	Wrapper	Hybrid
Dependency on data mining algorithm	Uses intrinsic characteristic of data	Uses data mining algorithm	Uses mining algorithm with intrinsic properties
Computational complexity	Low	High	High
Accuracy of result	Low	High	High
Suitability towards dimension of dataset	Very High	High	High

Feature Selection: Filter Methods

- In these methods, feature evaluation function use intrinsic characteristics of the data set to select feature subsets by typically ranking individuals without engaging any data mining algorithms.
- some of the criteria used for evaluation are interclass distances, information or uncertainty, dependency, and consistency.

Summary of evaluation criteria for filter methods

Criteria	Key concept	Generality	Computational complexity	example
Distance	Based on separability of Instances	Yes	Low	Euclidian distance
Information	Based on information gain or Entropy reduction	Yes	Low	Information Gain
Dependency	Based on the values of correlation coefficient of two features	Yes	Low	Correlation coefficient
Consistency	Based on consistency of features toward class label	Yes	Moderate	Inconsistency rate

Feature Selection: Filter Methods (Contd.)

Algorithm: selection of relevant subset of features using filter method.

Input:

$DS(f_1, f_2, \ldots, f_n)$ is DataSet (DS) with attributes $f_1, f_2, \ldots, f_n$.

S_0 is an initial search subset.

T is termination criteria.

Output:

S_{opt} is an optimal subset. It will be obtained in the last iteration.

Algorithm:

Step 1: Initialize $S_{opt} = S_0$

Step 2: $Temp_{opt}$ = Evaluation of S_0 by an independent criterion

Step 3: Repeat 4 to 6 until termination criteria meet

Step 4: Generate a subset S for evaluation

Step 5: Temp = Evaluation of current subset S by independent criteria

Step 6: If (Temp is better than $Temp_{opt}$

{ $Temp_{opt}$ = Temp

S_{opt} = S

}

Feature Selection: Filter Methods (Contd.)

- Commonly used methods for feature selection are
  - Information Gain(IG): Information measures typically determine the Information gain or reduction in entropy when the data set is split on a feature.
  - Correlation coefficient(CC): The correlation coefficient is a numerical way to quantify the relationship between two features.
  - Symmetric uncertainty (SU): Features are selected based on highest symmetric uncertainty values between the feature and target classes.

Feature selection: Filter Methods (Contd.)

Summary important methods for evaluation in filter method

Name	Key Concept	Mathematical formula	Merits	Demerits
Information gain (IG)	For higher IG, feature is more relevant	IG(D|X) = I(D)- I(D|X)	Robust to the effect of outlier and act as white box classifier	Splits of features are very sensitive to training data.
Correlation Coefficient (CC)	For higher value of CC, Features are more redundant	$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{n\sigma_a\sigma_b}$	It is easy to work out and easy to interpret	It measures only linear relationships between features.
Symmetric uncertainty (SU)	Feature having higher value with target class is better	$SU(X_k, w) = 2\left[\frac{I(X_k, w)}{H(X_k) + H(w)}\right]$	Appropriate for high dimensional data set	It fails to take into consideration the interaction between features

Where IG of an attribute (X) is measured by the reduction in entropy (I). In CC, N is the number of instances, and are respective mean values of features a and b. In SU, I is mutual information of k^{th} feature and H is entropy of feature X.

Feature Selection: Wrapper Methods

- Wrapper methods require a predetermined mining algorithm for evaluating generated subsets of features of data set. It usually gives superior performance as they find features better suited to the predetermined mining algorithm .
- Within the Wrapper category, Predictive Accuracy is used for Classification, and Cluster Goodness for Clustering
- Some well-known classifiers whose training and testing are relatively fast are as follows:
  - K-nearest neighbor classifier
  - Linear Discriminant Analysis
  - Bayesian classifiers
  - Support Vector Machines(SVM)

Feature Selection: Wrapper Methods (Contd.)

- K-nearest neighbor (K-NN): This classical classifier classifies testing sample by assigning it to the class most frequently represented among the k nearest training samples. Neighborhood is found based on a distance metric such as Euclidian distance.
- Linear Discriminant Analysis (LDA): It seeks a projection that best separate the data. It maximizes the ratio of between-class (inter) variance to the within-class (intra) variance in any particular data set thereby guaranteeing maximal seperability.

Feature Selection: Wrapper Methods (Contd.)

- Support Vector Machine (SVM): SVM is a method for classification of both linear and non-linear data. It uses a nonlinear mapping to transform the original training data into higher dimension .Within new dimension, it searches for the linear optimal decision boundary for separating one class from another. SVM finds this support boundary (hyperplane) using support vectors (essential training instances).
- Bayesian Classifier :The Bayesian Classifier is a statistical classifier, which has the ability to predict the probability that a given instant belongs to particular class. In theory, Bayesian classifiers have the minimum error rate with respect to other classifiers. However, these classifiers are not suitable for high dimensional features space.

Feature Selection: Wrapper Methods (Contd.)

A brief review for literatures used for wrapper methods

Paper	Classifier/evaluator for clustering	Search method used
[1]	K-nearest neighbour	binary particle swarm optimization
[3]	K-nearest neighbour	Genetic Algorithm
[4]	Support Vector Machine	Particle swarm optimization
[5]	K-means clustering algorithm	Evolutionary local search algorithm (ELSA)

Feature Selection: Hybrid Methods

- Hybrid methods are approaches to take advantages of other two methods. These methods are more suitable for feature selection especially on high dimensional data sets.
- These methods include both the intrinsic characteristic of data set and predefined mining algorithm.

Feature Selection: Hybrid Methods (Contd.)

A brief review for literatures used for hybrid methods

Paper	Classifier/Evaluator for clustering	Filter method used	Search method used
[6]	K-nearest neighbor	Information Gain	Genetic algorithm
[2]	Neural network	Information Gain	Ant colony optimization
[7]	Neural Network	Correlation	Particle swarm optimization
[8]	K-nearest neighbor	Euclidian distance	Genetic algorithm

Feature selection: Stopping Criteria

- A stopping criterion determines when the feature selection process should stop.
- Some frequently used stopping criteria are:
  - The search completes.
  - Some given bound is reached, where a bound can be a specified number (minimum number of features or maximum number of iterations).
  - Subsequent addition (or deletion) of any feature does not produce a better subset .
  - A sufficiently good subset is selected (e.g., a subset may be sufficiently good if its classification error rate is less than the allowable error rate for a given task).

Feature Selection: Result Validation

- By comparing previously related produced results in literatures.
- A straightforward way for result validation is to directly measure the result using prior knowledge about the data.
- Knowledge on the irrelevant or redundant features can also help.
- Some indirect methods by monitoring the change of mining performance with the change of features.

Feature Extraction

- Feature extraction is process of extracting new set of reduced features from original features based on some attributes transformation.
- The transformed features are linear combinations of the original attributes.
- Some important approaches are: Principal Component Analysis (PCA), Neural networks approach (NN), Independent Component Analysis (ICA), Fisher Discriminate Analysis (FDA), and Multidimensional Scaling (MDS).

Feature Extraction (Contd.)

- PCA and ICA are two frequently used methods for feature extraction .
  - Principal Component Analysis (PCA) : PCA is a method in which x ($x < n$) new attributes are formed from the linear combination of n original attributes. Key steps for PCA are as follows.
    1. Calculate the covariance matrix for features.
    2. Calculate eigenvector and eigen values of the covariance matrix. Eigenvector with higher eigen value is the principal component of data set.
    3. Choose the components and form the new feature vectors.

Feature Extraction (Contd.)

- Independent Component Analysis (ICA)
  - ICA minimize the statistical dependence between the basis vectors.
  - Sometimes, PCA can be used as preprocessing step in some ICA algorithm.

Feature Extraction (Contd.)

Comparison of PCA and ICA

Principal component analysis (PCA)	Independent component analysis (ICA)
Goal of PCA is to minimize the reprojection error.	Goal of ICA is to minimize the statistical dependence between the basis vectors.
Basis vectors are orthogonal and ranked in order	Basis vector are neither orthogonal nor are in order.
Generally, basis vector are less expensive to compute.	Generally, basis vector are more expensive to compute.
Basis vectors are less spatially localized	Basis vectors are spatially more localized and statically independent.
It is relatively older method	It is relatively newer method

Hybrid Methods

- The methods are two stage methods. They use a combination of feature selection – feature selection (FS-FS) methods or feature selection – feature extraction (FS-FE) methods. The FE-FS combination is not considered as a good approach as FE methods simply transform the dataset to lower dimensions retaining all the characteristics of the original dataset.
- Recently, a three-stage approach has also been proposed for dimension reduction and the authors claim that a combination of FS-FS-FE method is a good approach as it removes irrelevant dimensions, redundant dimensions and noisy dimensions, i.e., retains only meaningful dimensions.

Multi-dimensional Approach for Feature Selection

- As the two objectives of the feature selection (i) minimization of the number of features in selected subset and (ii) increasing predictive accuracy of classifier and cluster goodness for clustering, are contradicting to each other, solving feature selection task through multi-objective optimization methods is good alternative.
- MOGA have been successfully used in feature selection [7] to generate the set of alternative solutions.
- Non-dominated sorting genetic algorithm (NSGA) is also focused by researchers to search for optimal set of solutions with the above two objectives minimizing the number of features used in classification and minimizing classification error [7].

Summary

- Feature selection is a fundamental approach to enhance the classification and clustering accuracy.
- Based on a brief literature survey, it may be concluded that two-stage / three-stage data reduction hybrid methods are more suitable in order to speed up the clustering and classification task as well as for enhancing accuracy of classifier and goodness of cluster respectively.

References

[1] L. Y. Chuang, C. H. Yang and J. C. Li, “Chaotic maps based on binary swarm optimization for feature selection”, Applied Soft Computing, 11, 239-248, 2011

[2] M.M. Kabir, M. Shahjahan, K. Murase. A New hybrid Ant colony optimization algorithm for feature selection. Expert system with applications.39, 3747-3763, 2012

[3] M. S. Sainin and R. Alfred, “A Genetic Based Wrapper Feature Selection Approach Using Nearest Neighbour Distance Matrix”, 3rd Conference on Data Mining and Optimization (DMO), 28-29, June 2011.

[4] C. J. Tu, L. Y. chuang, J. Y. Chang and C. H. Yang, “ Feature selection using PSO-SVM”, In proceeding of multi-conference of engineers, 138-143, 2006.

References (Contd.)

[5] Y. S. Kim, W. N. Street and F. Menczer, “Feature selection in unsupervised learning via evolutionary search”, In Proceedings of ACM SIGKDD International Conference on Knowledge and Discovery, USA, 365–369, 2000

[6] C. H. Yang, L. Y. Chuang and C. H. Yang, “IG-GA, A hybrid filter/wrapper method for feature selection of microarray data”, Journal of Medical and Biological Engineering, 30(1), 23-28, 2009.

[7] C. Jin, S. W. Jin and L. N. Qin, “ Attribute selection method based on a hybrid BPNN and PSO algorithms”, Applied soft computing, 12, 2147-2155, 2012.

[8] L. Y. Chang, K. C. Wu and C. H. Yang, “ Hybrid feature selection algorithm using gene hybrid data”, IEEE Conference on Soft Computing in Industrial Applications (SMCia/08), Kaohsiung, 25-27, June 2008.

Thank You

Pramod Kumar Singh
pksingh@iiitm.ac.in | pksingh7@gmail.com
+91 94 25 773268