SPPS: Supervised Projected Clustering Method Based on Particle Swarm Optimization

Satish Gajawada and Durga Toshniwal

***Abstract*— Supervised clustering algorithms are applied on classified examples with the goal of determining class-uniform clusters. These methods evaluate clustering solutions based on class impurity unlike traditional clustering methods. These methods can be used for tasks like data editing and learning of subclasses to enhance classification methods. Supervised clustering methods have been proposed in literature to find class-uniform full dimensional clusters. But for high dimensional dataset with subspace clusters there is need for supervised clustering method which finds class-uniform subspace clusters. In this paper we propose Supervised Projected clustering Particle Swarm optimization method (SPPS method). The proposed method has been applied on Wisconsin breast cancer data to find subspace clusters present in this dataset. The SPPS method may be used for pre-processing of high dimensional datasets with subspace clusters.**

***Index Terms*—Supervised clustering, projected clustering, particle swarm optimization, pre-processing.**

I. Introduction

Most clustering methods are unsupervised clustering methods. Supervised clustering methods are applied on classified data. The goal of supervised clustering algorithms is to obtain clusters that are pure in terms of class distribution. The fitness function of supervised clustering method is based on class purity. Supervised clustering can be used for data editing. Subclasses can be learnt and used for enhancing classification methods using supervised clustering. Semi-supervised clustering has some similarity with supervised clustering. The goal of semi-supervised clustering is to use little side information like small set of classified examples available to enhance clustering algorithm. The fitness function of semi-supervised clustering algorithm is based on class purity and objective functions of traditional clustering algorithms [1].

Subspace and projected clustering methods find clusters that exist in subspaces of dataset. In subspace clustering one point may belong to more than one subspace cluster. Hence projected clustering is preferred over subspace clustering when partition of points is required. Particle swarm optimization is well known for solving optimization problems. Particle swarm optimization (PSO) was applied in literature for soft projected clustering. Recently, PSO has been applied for hard projected clustering [2].

The structure of dataset obtained using clustering method can be used to pre-process the dataset to obtain better classifiers. Recently, a hybrid method for classification is proposed using projected clustering in the pre-processing step. When the dataset contains subspace clusters it is appropriate to use clustering method that finds subspace clusters in the preprocessing step to classification [3].

Various PSO based unsupervised clustering methods have been proposed in literature. Initial seeds to k-means clustering may be given based on result obtained from PSO based clustering method [4]. K-means clustering may be used to seed the initial swarm of PSO [5]. But these methods may fail to find clusters that exist in subspaces of the dataset. Hence PSO has been applied to find subspace clusters. In PSOVW, PSO solves variable weighting problem in soft projected clustering [6]. PSO is applied to find optimal cluster centers of subspace clusters by optimizing a subspace cluster validation index [2].

Clustering results are evaluated by using cluster validation indices. The information present in the data is used for performing internal cluster validation. External information related to data is used for external cluster validation [7]. Many clustering methods need the number of clusters to be provided as an input parameter. The optimal clustering solution can be known by executing clustering algorithm several times by changing the number of clusters. The partition at which cluster validation index gives best value is selected as optimal clustering solution [8].

The impurity of certain split in decision trees has been determined by using various impurity measures like gini index, entropy index and classification error index [9]. These impurity measures can be used for external cluster validation because clustering solution of a dataset can be viewed as a partition at a particular node in decision tree and various measures like gini index can be used to determine impurity of such partition [10].

In this paper, we propose a supervised projected clustering method based on PSO where optimal cluster centers of subspace clusters are found by using an external cluster validation index. The proposed method has been applied on Wisconsin breast cancer data from UCI Machine learning repository [16].

The remainder of paper is organized as follows: Related work is discussed in Section II. Proposed work is explained in Section III. Experimental results are given in Section IV. Section V gives conclusion and future work.

II. Related Work

Demiriz et al. [10] used hybrid DB-Gini index for semi-supervised clustering. But by making one of the two

Manuscript received April 5, 2012; revised April 30, 2012.

Authors are with the Indian Institute of Technology Roorkee, Roorkee, India (email: gajawadasatish@gmail.com; durgafec@iitr.ernet.in).

regularization parameters equal to zero the algorithm can me made supervised clustering algorithm. Eick et al. [1] introduced four algorithms for supervised clustering. Experimental results were presented to illustrate the benefits of supervised clustering for creating dataset summaries and to enhance existing classification methods.

The above methods find clusters that exist in full dimensional space. Various methods have been proposed in literature to find clusters that exist in subspaces of the dataset. In DOC a projected cluster contains at least a fraction α of the total number of points. All the attributes for which projection of points in projected cluster are contained in a segment of length w are taken as relevant attributes of projected cluster [11]. PROCLUS finds subspace clusters by selecting neighborhood for each medoid and identifying relevant attributes of subspace cluster based on the selected neighborhood [12]. HARP is a hierarchical clustering method. Relevance score is calculated for attributes and quality of a cluster is taken as the sum of the relevance scores of its relevant attributes. Two clusters are merged if the resulting cluster has d_{min} or more relevant attributes. Those attributes of merged cluster for which relevance score is greater than R_{min} are selected as relevant attributes. HARP requires percentage of outliers as input parameter [13]. A specialized distance measure and a full dimensional density based clustering algorithm are used in PreDeCon. Each point contains a separate weight vector for all attributes. A weight of $k>>1$ is received by those attributes for which the variance of the points in a full-dimensional ε-neighborhood of the point is smaller than a threshold δ. Remaining attributes receive weight 1. The relevant attributes for ε-neighborhood of the point are those attributes for which received weight k [14].

Lu et al. [6] proposed PSOVW for soft projected clustering of high-dimensional data. The problem of text clustering was handled by extending PSOVW in [15]. Satish Gajawada et al. [2] proposed PCPSO for finding optimal cluster centers of subspace clusters. Subspace clusters can be found by using optimal cluster centers given by PCPSO.

But all the clustering methods described above which finds subspace clusters are not supervised projected clustering methods. Although supervised full dimensional clustering methods have been proposed in literature but supervised projected clustering methods for high dimensional datasets with subspace clusters are not yet explored.

The combination of clustering and classification methods can yield better results compared to applying classification methods only. Several hybrid clustering and classification methods were proposed in the literature. The clustering result obtained in the pre-processing step can be used for various kinds of pre-processing steps. Clustering was used to add meta-features in the pre-processing step. Classification method was applied on the dataset with meta-features to get a better classifier [19]. Fang et al. [20] proposed a hybrid method using Naive Bayes method. The proposed method yielded better results compared to Naive Bayes method. Classification methods like SVM cannot be applied when the dataset is unlabelled. Clustering may be used to obtain class labels so that classification methods can be used to build the classifier on the dataset labelled by clustering. Maokuan et al. [21] proposed a classification method using support vector machines and K-means for classification of unlabeled data. K-means was used to assign class labels in the pre-processing step to classification stage.

A clustering and classification framework was proposed in [22] by using K-means for verification of valid grouping. K-means was used in [22] to get clusters. The points which were misclassified in clusters were deleted in the pre-processing stage. A better classifier is built by using pre-processed data. This clustering and classification framework was applied on several datasets from UCI Machine learning repository and it was observed that this hybrid framework for classification obtained promising classification accuracy compared to other methods found in literature.

All the hybrid clustering and classification methods described above used full dimensional clustering method in the pre-processing step. But when different clusters exist in different subspaces of dataset then there is need to use clustering method which finds subspace clusters in the pre-processing step to classification. Recently, Satish Gajawada et al. [3] proposed various hybrids projected clustering and classification methods. When the amount of available labelled data is very less than building the classifier using available limited labelled data may not yield good results. Hence PCPSO-Classification method was proposed in [3] to solve the problem of limited labelled high dimensional data. Different classification methods can be obtained by using existing classification methods in the classification stage of PCPSO-Classification method. The proposed methods were applied on datasets with limited labels to get better classification accuracy compared to applying classification methods directly without using PCPSO.

There is scope for using other projected clustering methods like SPPS method proposed in this paper for pre-processing high dimensional dataset to get a better classifier model.

III. Proposed Method

Subsection A explains some impurity measures which can be used in the proposed SPPS method for external cluster validation. Fig. 1 shows PCPSO method proposed in [2] for finding subspace clusters. Fig. 2 shows PCPSO-Classification method proposed in [3]. Fig. 3 shows proposed SPPS method. Subsection B explains proposed method.

A. External Cluster Validation Indices

Impurity measures like entropy index used in decision trees can be used for external cluster validation. In this section we show various impurity measures which may be used in SPPS method for external cluster validation. In Equation (1) to Equation (6), $p(i|t)$ represents fraction of points belonging to class i at node t, c represents number of classes that are present in the dataset and k represents the number of clusters in clustering solution. Equation (1) and Equation (2) gives gini and entropy measures of a cluster. Equation (3) and Equation (4) gives gain measures related to gini and entopy impurity measures. A single cluster containing all points in the dataset is viewed as parent node. The clusters obtained by using clustering method are viewed as child nodes. The gain measures in Equation (3) and Equation (4) gives gain obtained by splitting parent node into child nodes which clusters are given by clustering method. $N(child_j)$ and N represent number of points in the child node j and number of points in dataset respectively. Split information of a clustering solution is given by Equation (5). Equation (6) gives information gain ratio impurity measure.

$$\text{Entropy}(t) = -\sum_{i=1}^{c} p(i \mid t)\log p(i \mid t) \quad (1)$$

$$\text{Gini}(t) = 1 - \sum_{i=1}^{c} [p(i \mid t)]^2 \quad (2)$$

$$\text{Information Gain} = \text{Entropy}(\text{parent}) - \sum_{j=1}^{k} \frac{N(\text{child}_j)}{N}\text{Entropy}(\text{child}_j) \quad (3)$$

$$\text{Gini Gain} = \text{Gini}(\text{parent}) - \sum_{j=1}^{k} \frac{N(\text{child}_j)}{N}\text{Gini}(\text{child}_j) \quad (4)$$

$$\text{Split Information} = -\sum_{i=1}^{k} \frac{N(\text{child}_i)}{N}\log\frac{N(\text{child}_i)}{N} \quad (5)$$

$$\text{Information Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}} \quad (6)$$

B. *Description of Proposed SPPS Method*

Supervised Projected clustering Particle Swarm optimization method (SPPS method) obtains optimal centers of subspace clusters by optimizing an external cluster validation index. Projected Clustering Particle Swarm Optimization method (PCPSO method) is shown in Fig. 1 does not use external information available while finding subspace clusters. But in SPPS method subspace clusters are obtained by using available external information. Various impurity measures like gini index can be used for external cluster validation in SPPS method.

In SPPS method, PSO is used for getting optimal solution. Each particle is of length equal to number of subspace clusters which is supplied as an input parameter. Decoding the particle gives K points from the dataset which are selected as centers of K subspace clusters.

Subspace clusters are obtained in SPPS method by finding neighborhood of cluster centers, identifying relevant attributes based on neighborhood and assigning points to centers using relevant attributes found. Relevant attributes are found again and points are reassigned to cluster centers to get subspace clusters.

Subspace clusters obtained are validated using an impurity measure. Class labels are used to validate the subspace clusters in the fitness function. Equation (1) to Equation (6) shows external cluster validation indices that can be used to obtain fitness of particles in SPPS method. In this paper, Gini Gain measure shown in Equation (4) is used for external cluster validation.

The velocities of particles are calculated and positions of particles are updated in each iteration. The SPPS method returns optimal centers of subspace clusters after reaching the termination condition. The optimal subspace cluster centers are used for finding subspace clusters.

Fig. 2 shows PCPSO-Classification. In this method PCPSO is used to solve the problem of limited labeled high dimensional data. Similarly, the SPPS method may be used in the pre-processing step to classification stage.

```
PCPSO Algorithm

Generate initial population

WITH IN Particle swarm optimization

    1. Determine subspace cluster centers

    2. Find neighbourhood of cluster centers

    3. Identify relevant attributes of clusters

    4. Assign points to cluster centers using relevant
       attributes found in above step

    5. Use clusters identified in above step to find
        relevant attributes of clusters

    6. Reassign points to cluster centers using relevant
       attributes identified in above step

    7. Subspace cluster validation

    8. Calculate pBest of each particle and gBest of
       Population

    9. Calculate particle velocity of each particle and
       update particle position of each particle

END LOOP

RETURN optimal subspace cluster centers
```

Fig. 1. PCPSO method

```
PCPSO-Classification Algorithm

Generate initial population

WITH IN Particle swarm optimization
    1. Find subspace clusters
    2. Subspace cluster validation
    3. Calculate pBest of each particle and gBest of
       Population
    4. Calculate particle velocity of each particle and
       update particle position of each particle
END LOOP

RETURN optimal subspace cluster centers

Use labelled points and subspace clusters identified by PCPSO
to label the unlabelled points. Apply classification method on
pre-processed data.
```

Fig. 2. PCPSO-Classification method

IV. Experimental Results

We applied proposed SPPS method on Wisconsin breast cancer data from UCI Machine Learning Repository [16]. The Gini Gain impurity measure shown in Equation (4) has been used for external cluster validation. The Wisconsin breast cancer dataset has 699 objects with each object having 9 dimensions. The objects in this dataset with missing values are deleted before finding subspace clusters. We applied PCPSO and PCPSO-Classification methods on some synthetic datasets to show advantage of pre-processing using projected clustering for high dimensional datasets [17].

```
SPPS Algorithm

Generate initial population

WITH IN Particle swarm optimization

    1. Determine subspace cluster centers

    2. Find subspace clusters similar to method described
       in PCPSO method

    3. External cluster validation step:

        a) Calculate impurity measure of each subspace cluster
        b) Calculate impurity measure of clustering solution by
        using impurity measure of each subspace cluster
        c) Calculate gain measure of the clustering solution
        d) Calculate gain ratio measure of the clustering
        solution if gain ratio index is used for cluster validation

    4. Calculate pBest of each particle and gBest of
       Population

    5. Calculate particle velocity of each particle and
       update particle position of each particle

END LOOP

RETURN optimal subspace cluster centers
```

Fig. 3. The SPPS method

Fig. 4 shows the best fitness and mean fitness values of all the generations for Wisconsin breast cancer data. The fitness value of the best individual for that particular generation is known as best fitness value. The average of fitness values of all individuals for that particular generation is known as mean fitness value. The improvement of fitness values from generation to generation for Wisconsin breast cancer data can be observed from Fig. 4.

Subspace clusters that are present in the dataset are called as input clusters. Subspace clusters that are identified by using SPPS method are called as output clusters. The input clusters present in Wisconsin breast cancer data are represented with letters {A, B} and output clusters are represented with numbers {1, 2}.

TABLE I shows the matching points between input and output clusters for Wisconsin breast cancer data for 6 average number of dimensions per subspace cluster. The output cluster 1 matched to cluster B and there are 12 misclassified points in this subspace cluster. The output cluster 2 matched to input cluster A and there are 10 misclassified points in this subspace cluster. Hence 96.78% of points are correctly classified and 22 points are misclassified which can be observed from TABLE I.

TABLE I: MATCHING POINTS BETWEEN OUTPUT AND INPUT CLUSTERS OF WISCONSIN BREAST CANCER DATA FOR AVERAGE SUBSPACE DIMENSIONS 6

Cluster	A	B
1	12	229
2	432	10

In [18] SUBCAD, a method for clustering high dimensional categorical datasets was proposed. SUBCAD was applied on Wisconsin breast cancer data and 87.55% of points were correctly classified in [18].

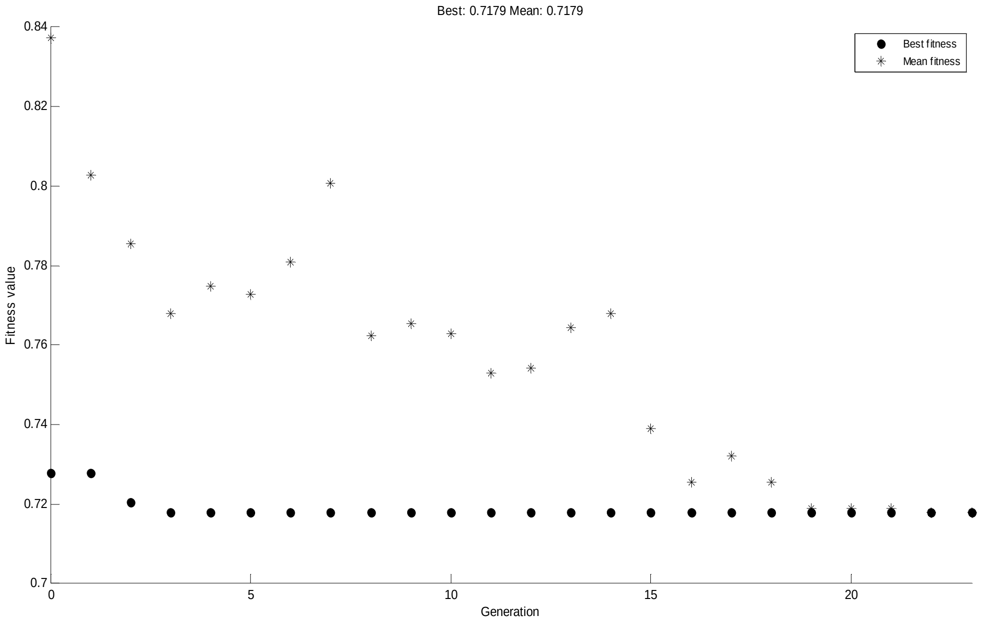

Fig. 4. Fitness values of Wisconsin breast cancer data for all generations

Table II shows results obtained by applying various classification methods on a synthetic dataset [17]. The synthetic dataset has 9 subspace clusters with 14 average number of relevant dimensions per subspace cluster. We have considered labels of less number of points in synthetic dataset to show the advantage of PCPSO-Classification compared to classification without using PCPSO in the pre-processing step for classification of high dimensional datasets with subspace clusters. Randomly 3 percent of points in synthetic dataset have been selected as labeled points. From Table II we can observe that directly performing classification with Decision table on synthetic dataset gave accuracy around 79 percent where as PCPSO-Decision table gave more than 91 percent accuracy. This significant difference in classification accuracy between PCPSO-Decision table and Decision table classification is due to that fact that the data has very limited labeled data. But the amount of available labels was increased by using PCPSO and limited labels present in the dataset. After pre-processing step the amount of labeled data is not limited and hence classification accuracy improved significantly. From Table II we can find that PCPSO-Naive bayes, PCPSO-Multi layer perceptron showed improvement compared to Naive bayes and Multi layer perceptron respectively.

TABLE II: CLASSIFICATION ACCURACY OBTAINED BY USING DIFFERENT CLASSIFIERS ON SYNTHETIC DATASET

Classification method	Accuracy
Naive bayes	84.8093
Multi layer perceptron	90.4977
Decision table	78.9916
PCPSO-Naive bayes	91.2853
PCPSO-Multi layer perceptron	96.1755
PCPSO-Decision table	91.8495

External validation indices used in this paper measures the quality using only available information in the form of class labels. But each subspace cluster is associated with set of relevant dimensions in addition to set of points. Hence there is scope to use cluster validation measures which measures the quality of clusters using relevant dimensions of subspace clusters as well.

V. CONCLUSION

In this paper we proposed Supervised Projected clustering Particle Swarm optimization method (SPPS method). In SPPS method, PSO obtains optimal cluster centers of subspace clusters by optimizing an external cluster validation index. The SPPS method has been applied on Wisconsin breast cancer data to find subspace clusters and it has been observed that 96.78% of points have been correctly classified.

We have also applied PCPSO-Classification method on synthetic dataset and showed that pre-processing the high dimensional data using projected clustering can improve classification accuracy significantly. PCPSO-Classification method solves the problem of limited labeled high dimensional data using subspace clusters in the pre-processing step to classification method.

Our future work includes creation of new methods for classification using supervised projected clustering methods. There is scope for creation of new supervised projected clustering methods using other optimization methods like Differential Evolution (DE) similar to PSO based supervised projected clustering method proposed in this paper.

REFERENCES

[1] C. Eick, N. Zeidat and Z. Zhao, “Supervised clustering–algorithms and benefits,” in *International Conference on Tools with Artificial Intelligence*, 2004, pp. 774–776.

[2] Satish Gajawada and Durga Toshniwal, “Projected clustering using Particle swarm optimization,” in *International Conference on Computer, Communication, Control and Information Technology*, 2012.

[3] Satish Gajawada and Durga Toshniwal, “Projected clustering particle swarm optimization and classification,” in *International Conference on Machine Learning and Computing*, 2012.

[4] X. Cui, T. E. Potok and P. Palafhingal, “Document clustering using particle swarm optimization,” in *2005 IEEE Swarm Intelligence Symposium*, 2005, pp. 185-191.

[5] D. W. Van der Merwe, A. P. Engelhrecht, “Data clustering using particle swarm optimization,” in *2003 Congress on Evolutionary Computation*, 2003, pp. 215-220.

[6] Y. Lu, S. Wang, S. Li, and C. Zhou, “Particle swarm optimizer for variable weighting in clustering high-dimensional data,” *Mach. Learn. 2011*, vol. 82, pp. 43-70, 2011.

[7] Erendira Rendon, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz, “Internal versus external cluster validation indexes,” *International Journal of Computers and Communications,* vol. 5, no. 1, 2011.

[8] Nadia Bolshakova and Francisco Azuaje, “Machaon CVE: cluster validation for gene expression data,” *Bioinformatics,* vol. 19, no. 18, pp. 2494-2495, 2003.

[9] Pang Ning Tan, Michael Steinbach and Vipin Kumar, *Introduction to Data Mining*, Pearson education, 2009.

[10] A. Demiriz, K. P. Bennett and M. J. Embrechts, “Semi-supervised clustering using genetic algorithms,” in *Artificial neural networks in engineering,* 1999, pp. 1-20.

[11] C. M. Procopiuc, M. Jones, P. K. Agarwal and T. M. Murali, “A monte carlo algorithm for fast projective clustering,” in *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, 2002, pp. 418-427.

[12] C. C. Aggarwal, C. M. Procopiuc, J.L. Wolf, P.S. Yu and J.S. Park, “Fast algorithms for projected clustering,” in *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, 1999, pp. 61-72.

[13] K. Y. Yip, D. W. Cheung and M. K. Ng, “HARP: A practical projected clustering algorithm,” *IEEE Trans. on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1387-1397, 2004.

[14] C. Bohm, K. Kailing, H. P. Kriegel and P. Kroger, “Density connected clustering with local subspace preferences,” in *Proceedings of the 4th International Conference on Data Mining (ICDM)*, 2004, pp. 27-34.

[15] Y. Lu, S. Wang, S. Li and C. Zhou, “Text Clustering via Particle Swarm Optimization,” in *IEEE Swarm Intelligence Symposium*, 2009, pp. 45-51.

[16] A. Frank and A. Asuncion, UCI Machine Learning Repository. [http://archive. ics.uci. edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[17] E. Muller, S. Gunnemann, I. Assent, and T. Seidl, “Evaluating clustering in subspace projections of high dimensional data,” in *Proc. 35th International Conference on Very Large Data Bases (VLDB)*, 2009.

[18] G. Gan and J. Wu, “Subspace clustering for high dimensional cate-gorical data,” in *ACM SIGKDD E xplorations News letter*, vol. 6, no. 2, pp. 87–94, 2004.

[19] A. Kyriakopoulou and T. Kalamboukis, "Using clustering to enhance text classification," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 805-806.

[20] Y. C. Fang, S. Parthasarathy and F. Schwartz, “Using clustering to boost text classification,” in *Proceedings of the IEEE ICDM Workshop on Text Mining*, 2002.

[21] Li Maokuan, Cheng Yusheng and Zhao Honghai, “Unlabeled data classification via Support Vector Machines and k-means clustering,” in *Proceedings of the International Conference on Computer Graphics, Imaging and Visualization*, 2004.

[22] B. M. Patil, R. C. Joshi, and Durga Toshniwal, "Effective framework for prediction of disease outcome using medical datasets: clustering and classification," *Int. J. Computational Intelligence Studies*, vol. 1, no. 3, pp. 273-290, 2010.

Satish Gajawada is presently an Undergraduate student of Computer Science and Engineering at the Indian Institute of Technology Roorkee, Roorkee, India.

His areas of interest include data mining and soft computing techniques. He has published his research work in various international conferences of repute.

Dr. Durga Toshniwal is presently working as an Assistant Professor at the Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, India. She completed her Doctor of Philosophy from Indian Institute of Technology Roorkee, India. Previously she worked as a Software Consultant in USA for some years and then pursued her research in data mining. Some of her areas of research interests include – time series data mining, web mining, privacy preserving data mining, data stream mining, applying soft computing techniques in data mining and text mining. Dr. Durga has published her research work in various international journals and conferences. She has attended, chaired sessions and presented her work in several reputed international conferences in USA, Australia, and Europe.