GA\textit{SSC: GENETIC ALGORITHM BASED SEMI-SUPERVISED CLASSIFIER}

A.S.Chandel, Suneet Kumar Gupta, Ashok Bhansali
Department of Computer Science & Engineering
O.P.Jindal Institute of Technology, Raigaru, India 496 109
Arvind.chandel@opjit.edu.in, suneet.gupta@opjit.edu.in, ashok.bhansali@opjit.edu.in

Abstract

In this paper, a Genetic algorithm based Semi-Supervised Classification Algorithm (GA-SSC) is proposed to classify the Semi-labeled data. Algorithm starts with clustering of semi-labeled data. For clustering two-stage genetic algorithm approach has been used. After Clustering of the semi-labeled data, labeling of cluster will be performed. In labeling process each cluster will be assigned a class to classify multi-class semi-labeled data. Labeling of cluster have been done using majority voting concept. The proposed algorithm is exhaustively tested with different benchmark datasets and results shows GASSC derived better performance on the search of cluster number and higher accuracy on classification problem.

Keywords— Semi-Supervised Classification, Genetic Algorithm, Two stage selection, Mutation.

I. INTRODUCTION

Machine learning algorithms depends on the availability of enough labeled data, however labeled data are hard to obtain in many application domains, while unlabeled data are easily available in abundance. As a consequence, semi-supervised learning has attracted much attention in recent years in many different fields ranging from bioinformatics to Web mining. With SSC we may pursue two different objectives: transductive and inductive classification [1]. The former is devoted to predicting the correct labels of a set of unlabeled examples that is also used during the training phase. The latter refers to the problem of predicting unseen data by learning from labeled and unlabeled data as training examples.

A variety of SSL algorithms have been proposed Existing SSC algorithms are usually classified depending on the conjectures they make about the relation of labeled and unlabeled data distributions. Broadly speaking, they are based on the manifold and/or cluster assumption. The manifold assumption is satisfied if data lie approximately on a manifold of lower dimensionality than the input space [2]. The cluster assumption states that similar examples should have the same label. Graph-based models [3] are the most common approaches to implementing the manifold assumption. As regards examples of models based on the cluster assumption, we can find generative models [3] or semi-supervised support vector machines [4]. Recent studies have addressed multiple assumptions in one model.

Thomas and Tiwari[5] proposes a SVM based semi supervised classifier approach to classify semi labeled data. CS-SCA by Chandel and Tiwari [6] are constructive semi supervised classification approaches to improve the performance of classifier. CS-SCA cannot handle the problem of confusing sample. Data sample which is not covered in learning process are confusing sample.

In this paper, we propose a genetic algorithm based novel approach to enhance performance of CS-SCA[6] classifier. This approach has two main steps, first is clustering of dataset by using two stage genetic algorithms and second is labeling of cluster. The advantage of GASSC can be illustrated by using it in classification problem. GASSC also eliminates confusing sample problem of CS-SCA. As GASSC is a semi-supervised approach, it requires less human effort. GASSC approach is tested with number of benchmark datasets and compared with CS-SCA and SVM based classifier.

The paper is organized as follows. Section 2 provides description of two stage genetic algorithm. Section 3 explains the method of labeling of clusters. Experimental results to demonstrate the usefulness of our approach have presented in section 4. Concluding remarks are given in section 5.

II. TWO STAGE GENETIC CLUSTERING

Suppose the clustering involves the partitioning of a set $A$ of objects into a collection of mutually disjoint subsets $C_i$ of $A$. Formally, let us consider a set of $M$ objects $A = \{a_1,a_2,\ldots,a_M\}$ to be clustered, where each object $a_i$ has $P$ attributes, i.e., $a_i = \{v_1, v_2, \ldots, v_P\}$. Thus, $A \subseteq \mathbb{R}^P$, and $M$ objects must be clustered into non-overlapping groups $C = \{C_1, C_2, \ldots, C_k\}$, where $k$ is the number of clusters, $C = \{C_1, C_2, \ldots, C_k\}$, and $C_i \cap C_j = \emptyset$, for $i \neq j$. The centers of each cluster form the matrix $Z=[z_1, z_2, \ldots, z_k]$ and the object number in each cluster form an array $B=[b_1, b_2, \ldots, b_k]$. According to [10], if $k$ is unknown, its value can be any integer in the set $K$, where $K = [k_{min}, k_{max}]$ and $k_{min} = 2$ and $(\ ) \ max k = \ round M$. To find the best partition pattern $C_k$ of $A$ is to search for the best number $k_0$ of clusters in $K$ and the optimal values $Z_o$ of cluster centers in $S$. 

ISSN No: 2250-3536 Volume 3, Issue 3, May 2013 151
A. **Chromosome representation**

We implemented variable length chromosomes with the cluster-center-based representation technique. Suppose there are \( N \) individuals in the initial population. Each individual \( \mathbf{X}_i \) represents one partition pattern of the data set, i.e. a certain value \( k \) of \( k \) and one set of cluster centers \( \mathbf{Z}_i \). Since \( k_i \) varies for different individuals, the length of \( \mathbf{X}_i \) is variable, where \( k_i \in K \). Hence, if each object has \( P \) real-coded attributes, \( \mathbf{X}_i \) is a real code chromosome with variable length \( L_i \), and \( L_i = P \times k_i \).

B. **Initialization of Population**

It is known that the appropriate initial values of cluster centers greatly affect the quality of partitional clustering [7]. Generally, there are four categories for clustering initialization approaches, i.e. random sampling methods, distance optimization methods, density estimation methods and attribute feature methods. In GASSC attribute feature method choose initial clusters in the light of attribute features of data. Comparing with other three categories, attribute feature methods absorb the inherent information of the data, consequently avoiding the blindness in the selection of seed clusters.

C. **Fitness Function:**

A fitness function must be defined for the evaluation of the chromosomes. One of the most commonly used fitness function for genetic clustering algorithm consists in minimizing the sum of squared Euclidean distances of the objects to their respective cluster centers. It is known that many clustering validity criteria have also been utilized as objective functions. Currently, there are several validity indices, including Calinski–Harabasz index, Davies–Bouldin index, Maulik–Bandyopadhyay index, Dunn index and Silhouette index. Recently Hruschka et al. proposed a simplified Silhouette index (SS index). We choose Calinski–Harabasz (CH) index as the fitness function of GASSC. The Calinski–Harabasz index is based on the internal cluster cohesion and the external cluster isolation. The corresponding internal cohesion is calculated by the within-group sum of square distances (WGSD) which is defined as

\[
\text{WGSD} = \sum_{i=1}^{N} \sum_{j \in k_i} (x_{ij} - \bar{x}_i)^2
\]

where \( x_{ij} \) is the value of the \( j \)-th element of the \( i \)-th object and \( \bar{x}_i \) is the average of \( x_{ij} \) for the \( i \)-th object. The external cluster isolation is calculated by the between-group sum of square distances (BGSD) which is defined as

\[
\text{BGSD} = \sum_{i=1}^{K} \sum_{j \in k_i} (x_{ij} - \bar{z}_i)^2
\]

where \( z_{ij} \) is the value of the \( j \)-th element of the \( i \)-th object in the cluster center and \( \bar{z}_i \) is the average of \( z_{ij} \) for the \( i \)-th object. The fitness function is defined as

\[
\text{Fitness}(\mathbf{Z}) = \frac{\text{WGSD}}{\text{BGSD}}
\]

D. **Genetic Operation**

**Selection**

The selection process plays an important role of focusing the search effort on promising regions in the data space and controlling the speed of convergence. In the clustering process, if the best number of clusters is unknown, \( Z_b \) can be obtained only when \( k_b \) is found. Therefore, the search of \( k_b \) should be the first essential task to be accomplished. For the selection process, we select some valid chromosomes with higher fitness value. The individuals with better fitness values have better chances of selection. There are several selection methods, such as Roulette-wheel selection, rank selection, tournament selection and soon. We use tournament selection in our method for selecting the chromosomes with best fitness values from the population. The selected chromosomes are applied to produce new child chromosomes (offspring) by the crossover and mutation operation as described in the following section.

**Parallel Crossover**

Crossover is a probabilistic process that exchanges information between a pair of parents to generate two offsprings. In the GASSC, the individuals with different numbers of clusters represent different data partitions. The crossover operation is significant only when the individuals are crossed under the same number of clusters. Therefore, the individuals in the subpopulation with same \( k \) undergo crossover, while the individuals corresponding to different \( k \)'s will not exchange genes. The parallel crossover is implemented in subpopulations under one-point crossover with a fixed crossover probability. Let \( \mathbf{X}_i \) and \( \mathbf{X}_j \) be two individuals in the same subpopulation. In the whole clustering process, only the first \( q \) genes of \( \mathbf{X}_i \) and \( \mathbf{X}_j \) will be exchanged and the other genes of \( \mathbf{X}_i \) and \( \mathbf{X}_j \) will be kept to the next generation. The value of \( q \) can be any integers in the region \([0, \min(L_i, L_j)]\), where \( L_i \) and \( L_j \) are the code lengths of \( \mathbf{X}_i \) and \( \mathbf{X}_j \), respectively.

**Mutation**

The intuition behind the mutation operation is to introduce some extra variability into the population to guarantee the global search in data space. Hence, in order to find the optimal partition in \( S \), the mutation probability \( p_m \) varies with and the variation tendency of \( p_m \) can be also divided into two stages. The mutation probability GASSC is given by

\[
p_m(i) = \begin{cases} 
0.1(1-k_{con}), & 0 < k_{con} \leq 0.9, \text{mutation in different } k_i's \\
\exp(-0.1/k_{con}), & 0.9 < k_{con} \leq 1, \text{mutation in different } Z_i's
\end{cases}
\]

In the first stage, i.e. when \( 0 < k_{con} < 0.9 \), the algorithm emphasize on the search of \( k_b \). \( p_m \) is the same for the whole population and decreases as \( k_{con} \) increases. Furthermore, the mutation operation only acts on \( k_i \). Note that \( k_i \) of \( \mathbf{X}_i \) is mutated as \( k_i' \) where \( k_i' \) can be a random integer value in \( K \). In the second stage, i.e. when \( 0.9 < k_{con} < 1 \), the search will naturally transfer to \( Z_b \) from \( k_b \). The numbers of clusters in the population gradually tend to be identical. The mutation does not work on \( k_i \), but on cluster centers of the individual. Additionally, in terms of the quality of clustering, the mutation probability of each individual will be reduced as its fitness increases. The individuals suffer from the bi-direction mutation. Let \( x_{ij} \) \((i=1,2,...,L_i)\) be the value on the \( j^{th} \) gene bit of \( \mathbf{X}_i \) \((i=1,2,...,N)\), and let \( x_{ij}' \) be the gene value of \( x_{ij} \)
after bi-direction mutation. Then with mutation probability \( p_i (m) \), \( x_{ij} \) is given by
\[
  x_{ij} = \begin{cases} 
    x_{ij} + T_{ij}(x_{j\text{max}} - x_{ij}), & T_{ij} \geq 0 \\
    x_{ij} + T_{ij}(x_{j\text{min}} - x_{ij}), & T_{ij} < 0 
  \end{cases}
\]

III. Summary of GASSC

On the basis of the above discussion, the summary of the proposed algorithm can be described as follows:

1. Initialization
   a) Input objects number \( M \), iteration number \( G_m \), crossover probability \( P_c \), evolutionary population \( N \), and array of the numbers of clusters, i.e. \( K = (k_1, k_2, \ldots, k_N) \).
   b) Adopt the maximum attribute range partition method to choose initial cluster centers for \( N \) times to form \( N \) initial individuals.

2. Evaluation of individuals
   a) Obtain new cluster centers by k-means;
   b) Cluster the objects according to new cluster centers and calculate the fitness of the individuals on the basis of Eq.(3);
   c) Keep the best individuals to the next generation for continuous evolution.

3. Generation of new individuals by genetic operations
   Each individual is subjected to the following conditions:
   a) Two-stage selection with probability \( P_s(i) \);
   b) Parallel cross over in the different subpopulations;
   c) Two-stage mutation with probability \( P_m(i) \).

4. Termination discrimination: when the number of iterations exceeds \( G_m \), stop the iteration and choose the best \( k_b \) and \( Z_b \) stored during iterative process as the result of algorithm. Otherwise, go to step 2 for iterating again.

5. Once the clustering is over, classification will be performed by using majority voting. For each cluster, count the number of sample related to each class, samples with particular class in majority will decide the class of that cluster.

IV. Conclusion

A genetic algorithm based methodology to classify the semi-labeled data has been proposed. This approach also handles the problem of confusing samples. Genetic Algorithm based Semi-supervised classifier is constructed for classifying the multi-class Semi-Labeled data as well as handing the problem of overlapping among clusters of different classes. The method allows the overlapping if it is among clusters of same classes. Then various benchmark datasets will be used to evaluate the performance of classifier (GASSC) in terms of accuracy, number of cluster and training time. It is expected that GASSC will give better classification accuracy and performance than of existing semi-supervised classifiers.

REFERENCES