An Optimized Clustering Algorithm Using Genetic Algorithm and Rough set Theory based on Kohonen self organizing map

1Asgarali Bouyer, 2Abdolreza Hatamlou
1Department of Computer Science
1Islamic Azad University – Miyandoab Branch
2Miyandoab, Iran
2University Kebangsaan Malaysia
2Selangor, Malaysia
1basgarali2@live.utm.my, 2hatamlou@iaukhoy.ac.ir

Abstract—The Kohonen self organizing map is an efficient tool in exploratory phase of data mining and pattern recognition. The SOM is a popular tool that maps high dimensional space into a small number of dimensions by placing similar elements close together, forming clusters. Recently, most of the researchers found that to take the uncertainty concerned in cluster analysis, using the crisp boundaries in some clustering operations is not necessary. In this paper, an optimized two-level clustering algorithm based on SOM which employs the rough set theory and genetic algorithm is proposed to defeat the uncertainty problem. The evaluation of proposed algorithm on our gathered poultry diseases data and iris data expresses more accurate compared with the crisp clustering methods and reduces the errors.

Index Terms- SOM, Clustering, Rough set theory, Genetic Algorithm.

I. INTRODUCTION

The self organizing map (SOM) proposed by Kohonen [1], has been widely used in industrial applications such as pattern recognition, biological modeling, data compression, signal processing and data mining [2]-[5]. It is an unsupervised and nonparametric neural network approach. The success of the SOM algorithm lies in its simplicity that makes it easy to understand, simulate and be used in many applications. The basic SOM consists of neurons usually arranged in a two-dimensional structure such that there are neighborhood relations among the neurons. After completion of training, each neuron is attached to a feature vector of the same dimension as input space. By assigning each input vector to the neuron with nearest feature vectors, the SOM is able to divide the input space into regions (clusters) with common nearest feature vectors. This process can be considered as performing vector quantization (VQ) [6]. Also, because of the neighborhood relation contributed by the interconnections among neurons, the SOM exhibits another important property of topology preservation.

Clustering algorithms attempt to organize unlabeled input vectors into clusters such that points within the cluster are more similar to each other than vectors belonging to different clusters [7]. The clustering methods are of five types: hierarchical clustering, partitioning clustering, density-based clustering, grid-based clustering and model-based clustering [8]. The rough set theory employs two upper and lower thresholds in the clustering process which result in a rough clusters appearance. This technique also could be defined in incremental order i.e. the number of clusters is not predefined by users.

Our goal is to optimized clustering algorithm that will use in poultry disease predictions. The clustering will assist in improving further analysis of the poultry symptoms data in detecting outliers. Analyzing outlier can reveal surprising facts hidden inside data like ambiguous patterns that are still assumed to belong to one of the predefined or undefined classes. Clustering is important in detecting outlier to avoid the high cost of misclassification. In order to cater for the complex nature of data of our problem domain, clustering technique based on machine learning approaches such as self organizing map (SOM), kernel machines, fuzzy methods, etc for clustering poultry symptoms (based on observation data – body, feathers, skin, head, muscle, lung, heart, intestines, ovary, etc) will prove to be a promising tool.

In this paper, a new two-level clustering algorithm is proposed. The idea is that the first level is to train the data by the SOM neural network and the clustering at the second level is a rough set based incremental clustering approach [9], which will be applied on the output of SOM and requires only a single neurons scan. The optimal number of clusters can be found by rough set theory which groups the given neurons into a set of overlapping clusters (clusters the mapped data respectively). Then the overlapped neurons will be assigned to the true clusters they belong to, by apply
genetic algorithm. A genetic algorithm has been adopted to minimize the uncertainty that comes from some clustering operations. In our previous work [3] the hybrid SOM and rough set has been applied to catch the involved ambiguity of clusters but the experiment results show that the proposed algorithm (Genetic Rough SOM) outperforms the previous one. The next important process is to collect poultry data from the common and important diseases, which can affect the respiratory and non-respiratory system of poultry. The first phase is to identify the format and values for input parameters from available information. The second phase is to investigate and develop data conversion and reduction algorithms for input parameters.

This paper is organized as following; in section II the basics of SOM algorithm are outlined. The basic of rough set incremental clustering approach are described in section III. In section IV the essence of genetic algorithm is described. The proposed algorithm is presented in section V. Section VI is dedicated to experiment results and section VII provides brief conclusion and future works.

II. SELF ORGANIZING MAP

Competitive learning is an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples in a specific domain of the input space. A division of neural nodes emerges in the network to represent different patterns of the inputs after training.

The division is enforced by competition among the neurons: when an input arrives, the neuron that is best able to represent it wins the competition and is allowed to learn it even better. If there exist an ordering between the neurons, i.e. the neurons are located on a discrete lattice, the competitive learning algorithm can be generalized. Not only the winning neuron but also its neighboring neurons on the lattice are allowed to learn, the whole effect is that the final map becomes an ordered map in the input space. This is the essence of the SOM algorithm. The SOM consist of \( m \) neurons located on a regular low-dimensional grid, usually one or two dimensional. The lattice of the grid is either hexagonal or rectangular.

The basic SOM algorithm is iterative. Each neuron \( i \) has a \( d \)-dimensional feature vector \( w_i = [w_{i1}, \ldots, w_{id}] \). At each training step \( t \), a sample data vector \( x(t) \) is randomly chosen for the training set. Distance between \( x(t) \) and all feature vectors are computed. The winning neuron, denoted by \( c \), is the neuron with the feature vector closest to \( x(t) \):

\[
e = \arg \min_i \| x(t) - w_i \| \quad i \in [1, \ldots, m]
\] (1)

A set of neighboring nodes of the winning node is denoted as \( N_c \). We define \( h_c(t) \) as the neighborhood kernel function around the winning neuron \( c \) at time \( t \). The neighborhood kernel function is a non-increasing function of time and of the distance of neuron \( i \) from the winning neuron \( c \). The kernel can be taken as a Gaussian function:

\[
h_c(t) = \frac{\| \text{pos}_i - \text{pos}_c \|}{\sigma(t)^2}
\] (2)

where \( \text{pos}_i \) is the coordinates of neuron \( i \) on the output grid and \( \sigma(t) \) is kernel width. The weight update rule in the sequential SOM algorithm can be written as:

\[
w_i(t+1) = \begin{cases} w_i(t) + \varepsilon(t) \sigma(t)(x(t) - w_i(t)) \forall i \in N_c, \\ w_i(t) \quad \text{otherwise} \end{cases}
\] (3)

Both learning rate \( \varepsilon(t) \) and neighborhood \( \sigma(t) \) decrease monotonically with time. During training, the SOM behaves like a flexible net that fold onto a cloud formed by training data. Because of the neighborhood relations, neighboring neurons are pulled to the same direction, and thus feature vectors of neighboring neurons resemble each other. There are many variants of the SOM [10, 11]. However, these variants are not considered in this paper because the proposed algorithm is based on SOM, but not a new variant of SOM.

The 2D map can be easily visualized and thus give people useful information about the input data. The usual way to display the cluster structure of the data is to use a distance matrix, such as U-matrix [12]. U-matrix method displays the SOM grid according to neighboring neurons. Clusters can be identified in low inter-neuron distances and borders are identified in high inter-neuron distances. Another method of visualizing cluster structure is to assign the input data to their nearest neurons. Some neurons then have no input data assigned to them. These neurons can be used as the border of clusters [13].

III. ROUGH SET INCREMENTAL CLUSTERING

This algorithm is a soft clustering method employing rough set theory [14]. It groups the given data set into a set of overlapping clusters. Each cluster is represented by a lower approximation and an upper approximation \( (A(C), \overline{A}(C)) \) for every cluster \( C \subseteq U \). Here \( U \) is a set of all objects under exploration. However, the lower and upper approximations of \( C \subseteq U \) are required to follow some of the basic rough set properties such as:

\[
\begin{align*}
(1) & \quad \emptyset \subseteq A(C) \subseteq \overline{A}(C) \subseteq U \\
(2) & \quad A(C) \cap \overline{A}(C) = \emptyset, \quad i \neq j \\
(3) & \quad A(C) \cap \overline{A}(C) = \emptyset, \quad i \neq j \\
(4) & \quad \text{If an object } u_k \in U \text{ is not part of any lower approximation, then it must belong to two or more upper approximations.}
\end{align*}
\]
Note that (1)-(4) are not independent. However enumerating them will be helpful in understanding the basic of rough set theory.

The lower approximation \( \underline{A}(C) \) contains all the patterns that definitely belong to the cluster \( C \) and the upper approximation \( \overline{A}(C) \) permits overlap. Since the upper approximation permits overlaps, each set of data points that are shared by a group of clusters defines indiscernible set. Thus, the ambiguity in assigning a pattern to a cluster is captured using the upper approximation. Employing rough set theory, the proposed clustering scheme generates soft clusters (clusters with permitted overlap in upper approximation).

For a rough set clustering scheme and given two objects \( u_i, u_k \in U \) we have three distinct possibilities:
- Both \( u_i \) and \( u_k \) are in the same lower approximation \( \underline{A}(C_i) \).
- Object \( u_i \) is in lower approximation \( \underline{A}(C) \) and \( u_k \) is in the corresponding upper approximation \( \overline{A}(C) \), and case 1 is not applicable.
- Both \( u_i \) and \( u_k \) are in the same upper approximation \( \overline{A}(C_i) \), and case 1 and 2 are not applicable.

The quality of a conventional clustering scheme is determined using within-group-error [15] \( \Delta \) given by:

\[
\Delta = \sum_{i=1}^{m} \sum_{u_i, u_k \in C_i} \text{distance} (u_i, u_k)
\]

where \( u_i, u_k \) are objects in the same cluster \( C_i \).

For the above rough set possibilities, three types of equation (4) could be defined as following:

\[
\begin{align*}
\Delta_1 &= \sum_{i=1}^{m} \sum_{u_i, u_k \in \underline{A}(C_i)} \text{distance} (u_i, u_k) \\
\Delta_2 &= \sum_{i=1}^{m} \sum_{u_i \in \underline{A}(X_i)} \sum_{u_k \in \overline{A}(X_i)} \text{distance} (u_i, u_k) \\
\Delta_3 &= \sum_{i=1}^{m} \sum_{u_i, u_k \in \overline{A}(X_i)} \text{distance} (u_i, u_k)
\end{align*}
\]

The total error of rough set clustering will then be a weighted sum of these errors:

\[
\Delta_{\text{total}} = w_1 \times \Delta_1 + w_2 \times \Delta_2 + w_3 \times \Delta_3 \quad \text{where} \quad w_1 > w_2 > w_3 \quad (6)
\]

Since \( \Delta_1 \) corresponds to situations where both objects definitely belong to the same cluster, the weight \( w_1 \) should have the highest value.

IV. GENETIC ALGORITHM

Genetic algorithm was proposed by John Holland in early 1970s, it applies some of natural evolution mechanism such as crossover, mutation, and survival of the fitness to optimization and machine learning. GA provides very efficient search method working on population, and has been applied to many problems of optimization and classification [16]-[17]. General GA process is as follows:

1. Initial the population of genes.
2. Calculates the fitness of each individual in the population.
3. Reproduce the individual selected to form a new population according to each individual’s fitness.
4. Perform crossover and mutation on the population.
5. Repeat step 2 through 4 until some condition is satisfied.

Crossover operation swaps some part of genetic bit string within parents. It emulates just as crossover of genes in real world that descendants are inherited characteristics from both parents. Mutation operation inverts some bits from whole bit string at very low rate. In real world we can see that some mutants come out rarely. Fig.1 shows the way of applying crossover and mutation operations to genetic algorithm. Each individual in the population evolves to getting higher fitness generation by generation.

\[
\begin{array}{c|c|c}
\text{Crossover} & \text{Mutation} \\
\hline
010000001001 & 011001011101 & 011001011101 \\
\hline
110010111101 & 111001000001 & 011100011001 \\
\hline
\end{array}
\]

Figure 1. Crossover and Mutation

V. GENETIC ROUGH SET CLUSTERING OF THE SELF ORGANIZING MAP

In this paper rectangular grid is used for the SOM. Before training process begins, the input data will be normalized. This will prevent one attribute from overpowering in clustering criterion. The normalization of the new pattern \( X_i = \{x_{i1}, \ldots, x_{id}\} \) for \( i = 1, 2, \ldots, N \) is as following:

\[
X_i = \frac{X_i}{\|X\|} 
\]

Once the training phase of the SOM neural network completed, the output grid of neurons which is now stable to network iteration, will be clustered by applying the rough set algorithm as described in the previous section. The similarity measure used for rough set clustering of neurons is Euclidean distance (the same used for training the SOM). In this proposed method
(see Fig. 2) some neurons, those never mapped any data are excluded from being processed by rough set algorithm.

From the rough set algorithm it can be observed that if two neurons are defined as indiscernible (those neurons in the upper approximation of two or more clusters), there is a certain level of similarity they have with respect to the clusters they belong to and that similarity relation has to be symmetric. Thus, the similarity measure must be symmetric.

According to the rough set clustering of the SOM, overlapped neurons and respectively overlapped data (those data in the upper approximation) are detected. In the experiments, to calculate errors and uncertainty, the previous equations will be applied to the results of SOM (clustered and overlapped data). Then for each overlapped neuron a gene is generated that represents the alternative distances from each cluster leader. Fig. 3 shows an example of the generated genes for \( m \) overlapped neurons on \( n \) existing cluster leaders.

\[
\begin{align*}
gene_1 & : d_1, d_2, d_3, \ldots, d_a, \ldots, d_n \\
gene_2 & : d_1, d_2, d_3, \ldots, d_a, \ldots, d_n \\
gene_3 & : \ldots \\
gene_m & : \ldots 
\end{align*}
\]

VI. EXPERIMENT RESULTS

To demonstrate the effectiveness of the proposed clustering algorithm GR-SOM (Genetic Rough set Incremental clustering of the SOM), two phases of experiments have been done on the well known Iris data set [18] and our gathered data. The Iris data set, which has been widely used in pattern classification, consists of 150 data points of four dimensions and our collected data has 48 data points. The Iris data are divided into three classes with 50 points each. The first class of Iris plant is linearly separable from the other two. The other two classes are overlapped to some extent.

The first phase of experiments, presents the uncertainty that comes from the data set and in the second phase the errors has been generated. The results of GR-SOM and RI-SOM [3] (Rough set Incremental SOM) are compared to I-SOM [4] (Incremental clustering of SOM). The input data are normalized such that the value of each datum in each dimension lies in [0,1].

For training, SOM 10×10 with 100 epochs on the input data is used. The general parameters for the genetic algorithm have been configured as Table I.

![Figure 2. Clustering of the Self Organizing Map. The overlapped neurons are highlighted for two clusters.](image)

![Figure 3. Generated genes. \( m \) number of overlapped neurons and \( n \) is number of existing clusters. The highlighted \( d_i \) is the optimize one that minimize the fitness function](image)

The aim of the proposed approach is making the genetic rough set clustering of the SOM to be as precise as possible. Therefore, a precision measure needs to be used for evaluating the quality of the proposed approach. A possible precision measure can be defined as the following equation [14]:

\[
\text{certainty} = \frac{\text{Number of objects in lower approx}}{\text{Total number of objects}}
\]

After the genes have been generated the genetic algorithm is employed to minimize the following fitness function which represents the total sum of each \( d_j \) of the related gene:

\[
F = \sum_{i=1}^{m} \sum_{j=1}^{n} g_i(d_j)
\]

TABLE I. GENERAL PARAMETERS OF THE GENETIC ALGORITHM

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Size</td>
<td>50</td>
</tr>
<tr>
<td>Number of Evaluation</td>
<td>10</td>
</tr>
<tr>
<td>Crossover Rate</td>
<td>0.25</td>
</tr>
<tr>
<td>Mutation Rate</td>
<td>0.001</td>
</tr>
<tr>
<td>Number of Generation</td>
<td>100</td>
</tr>
</tbody>
</table>

TABLE II. THE CERTAINTY-LEVEL OF GR-SOM, RI-SOM AND I-SOM ON THE IRIS DATA SET FROM EPOCH 100 TO 500.

<table>
<thead>
<tr>
<th>Epoch</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-SOM</td>
<td>33.33</td>
<td>65.23</td>
<td>76.01</td>
<td>89.47</td>
<td>92.01</td>
</tr>
<tr>
<td>RI-SOM</td>
<td>67.07</td>
<td>73.02</td>
<td>81.98</td>
<td>91.23</td>
<td>97.33</td>
</tr>
<tr>
<td>GR-SOM</td>
<td>69.45</td>
<td>74.34</td>
<td>83.67</td>
<td>94.49</td>
<td>98.01</td>
</tr>
</tbody>
</table>

http://sites.google.com/site/ijcsis/
ISSN 1947-5500
Figure 4. Comparison of the certainty-level of GR-SOM, RI-SOM and I-SOM on the Iris data set.

\[ \sum_{i=1}^{n} w_i = 1 \]

\[ \text{and for each } w_i \text{ we have:} \quad w_i = \frac{1}{6} (4-i). \] (10)

<table>
<thead>
<tr>
<th>Method</th>
<th>( \Delta_1 )</th>
<th>( \Delta_2 )</th>
<th>( \Delta_3 )</th>
<th>( \Delta_{\text{total}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris Data set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GR-SOM</td>
<td>1.05</td>
<td>0.85</td>
<td>0.04</td>
<td>1.4</td>
</tr>
<tr>
<td>I-SOM</td>
<td>2.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE III. COMPARATIVE GENERATED ERRORS OF GR-SOM AND I-SOM ON THE IRIS DATA SET ACCORDING TO EQUATIONS (5) AND (6).

Furthermore, to demonstrate the effectiveness of the proposed clustering algorithm (RI-SOM), two data sets, one artificial and one real word data set were used in our experiments. The results are compared to I-SOM (Incremental clustering of SOM). The input data are normalized such that the value of each datum in each dimension lies in \([0,1]\). For training SOM 10 \times 10 with 100 epochs on the input data is used.

![Graph of error comparison between I-SOM and RI-SOM](image)

The artificial data set is a 569 30-dimensional data set which is trained twice, once with I-SOM and once with RI-SOM. The errors of generated results are calculated from the difference between the results of equation (9) and 1, see “Fig. 5”.

From the “Fig. 5” it could be observed that the proposed RI-SOM algorithm generates less error in cluster prediction compare to I-SOM.

VII. CONCLUSION AND FUTURE WORK

In this paper a two-level based clustering approach (GR-SOM), has been proposed to predict clusters of high dimensional data and to detect the uncertainty that comes from the overlapping data. The approach is based on the rough set theory that employs a soft clustering which can detect overlapped data from the data set and makes clustering as precise as possible, then GA is applied to find the true cluster for each overlapped data. The results of the both phases indicate that GR-SOM is more accurate and generates fewer errors as compare to crisp clustering (I-SOM).

The proposed algorithm detects accurate overlapping clusters in clustering operations. As the future work, the overlapped data also could be assigned correctly to true clusters they belong to, by assigning fuzzy membership value to the indiscernible set of data. Also a weight can be assigned to the data’s dimension to improve the overall accuracy.

REFERENCES


