Parallel Coordinates for Visualization of Rules Developed using Grammatical Evolution

M. Abdulgader and D. Kaur

1 Department of Electrical Engineering and Computer Science, University of Toledo, Toledo, Ohio, USA

Abstract - Information visualization has become an increasingly interesting research area. Such visualization techniques are found to be useful in many applications. This paper illustrates the visualization of fuzzy rules generated by grammatical evolution using parallel coordinates. The application chosen was the well-studied machine learning problem of iris data and the fuzzy rules generated to classify it. The paper presents a novel technique using parallel coordinates to visually verify the correctness of each rule.

1. Introduction

The use of visualization techniques to view large scale multi-dimensional datasets is a well-known approach in data mining field. Visualization of data higher than three dimensions is not an easy and straight forward task using Cartesian coordinates. Representing n-dimensional dataset using Cartesian coordinates will require scatter plots taking two dimensional representation at time which makes the complexity of O(n^2). Inselberg [1] proposed the parallel coordinates method where the n-dimensional data is represented using n parallel lines. Therefore reducing the complexity of the n-dimensional data set to be O(n). With the N axes aligned as parallel lines; there is a correspondence of dimensions with their respective axis. The set of connected points on parallel axes is a polyline which represent one entry of the n-dimensional data set [1]. For illustrating the visualization with parallel coordinate, the Iris data by R.A. Fisher has been used. The three species of Iris flower have four input parameters which are: sepal width, sepal length, pedal width and pedal length [2].

Fig.1: Representation of Iris Data in Parallel Coordinates

2. Grammatical Evolution

Grammatical Evolution is a process of evolving a grammar template which is defined for each problem uniquely using Backus Naur Form (BNF) template [3, 4, 5]. For the Iris data the template was chosen in the form of fuzzy rules. The input parameters of Sepal Length (SL), Petal Length (PL), Sepal Width (SW) and Petal Width (PW) are uniformly divided into four fuzzy sets of Small, Small Medium, Medium and Large, as shown in Fig. 2. [2]

Fig.2: Triangular membership functions for Sepal-Length [2]

Input variable “Sepal-Length”

The representative tuple of BNF grammar is \{N, T, P S\}, where N represents an assortment of non-terminals, T is for the set of terminals, P shows the production rules which are necessary for the conversion of N to T, and S, an element of N, is the start symbol [6, 7]. Fig. 3 represents the BNF template for the fuzzy rules to be evolved for iris data [2].

The difference between using GE and other evolutionary approaches in the literature is that in GE the rule structure is evolvable within the constraint of the BNF. Also the transcription from bit string to evolved code uses redundancy to improve search efficiency[2,8]

Fig.3: BNF template for the fuzzy rules to be evolved
The following section of the BNF template defines the template of rules:

If (SL is <ant>) and (SW is <ant>) and (PL is <ant>) and (PW is <ant>) Then <cons>;
If (SL is <ant>) and (SW is <ant>) and (PL is <ant>) and (PW is <ant>) Then <cons>;
.
(repeat n times for n rules)
.
.
.
Each input parameter such as SL, PL SW and PW is associated with and <ant> which represents the antecedent to be evolved. The choices for <ant> are the fuzzy sets from Fig.3 and are shown below.

<ant> ::= Small (0)
| Small-Medium (1)
| Medium-Large (2)
| Large (4)
| All (5)

The options for <cons> are

<cons>::=
| Virginica (0)
| Setosa (1)
| Versicolour (2)

Eight fuzzy rules were obtained through the use of “Fuzzy Classification using Grammatical Evolution for Structure Identification” [2]. The classification success rate was 97.7%. Population size was 100 and it was randomly initialized, Number of generations was 50 and Crossover rate was 0.8.

Table 1. A Rule base on Fuzzy Classification using GE [2]

<table>
<thead>
<tr>
<th>Antecedent</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>If (SL is Large) and (PL is Large)</td>
</tr>
<tr>
<td>2</td>
<td>If (SL is Large) and (PL is Medium-Large) and (PW is Large)</td>
</tr>
<tr>
<td>3</td>
<td>If (SL is Small and (SW is Medium-Large) and (PL is Small)</td>
</tr>
<tr>
<td>4</td>
<td>If (PL is Small) and (PW is Small-Medium)</td>
</tr>
<tr>
<td>5</td>
<td>If (SL is Medium-Large) and (PL is Large) and (PW is Medium-Large)</td>
</tr>
<tr>
<td>6</td>
<td>If (SL is Small and (SW is Small-Medium) and (PL is Large) and (PW is Medium-Large)</td>
</tr>
<tr>
<td>7</td>
<td>If (PW is Medium-Large) and (PL is Medium-Large) and (PW is Large)</td>
</tr>
</tbody>
</table>

3. Two-dimensional Visualization

The conventional way of visualization of 4-dimensional data would be to generate scatter plots taking two parameters at a time. This would generate six plots as shown in Fig.4. From the scattered graphs if there is no overlap of colored dots (colors represent different categories of iris flower), it is easy to come up with rules using two antecedents which would classify the data correctly.

For instance from Fig. 4a it can be inferred that if Petal Length is small and Petal Width is small then the category is Setosa irrespective of sepal length and sepal width. In other words Sepal Length and Sepal Width are Don’t Care. Fig 4b and 3c also validate the same rule for category Setosa. This rule was again verified in parallel coordinates as shown in Fig. 4a.

However, there is a considerable overlapping among the categories of Virginica and Versicolour (among blue and magenta colors). Fig. 3b illustrates that if petal width is Small-Medium or Medium Large the category is always Versicolour. Fig. 4a illustrates that if Petal width is Small Medium and Petal Length is small medium the category is Versicolour. However from Fig. 4b it is also clear that if Petal Width is Medium Large and Sepal Width is Medium Large then the output category can be Versicolour or Virginica. In this case the boundaries are fuzzy and therefore no crisp rule can classify it correctly. Hence, in order to classify iris correctly we had evolved fuzzy rules.

Thus, it can be concluded that scatter plots are not enough to classify the overlapping iris data. The disadvantage of implementing two-dimensional representation obscures the efficient analysis of other parameters regarding their impact on the classification. Therefore it can be concluded that Cartesian representation is not suitable for classifying the overlapping data of iris flower.
4. Visualization of Rules in Parallel Coordinates

By applying the filters in the parallel coordinate representation of the iris data, figures 5a, 5b and 5c were generated. The data represented by burgundy are the instances of class Setosa. Fig. 5b shows the data which represent Versicolor in blue lines. The purple lines represent the data that belong to Virginica in Fig. 5c. Fig 5 represents the following three observations.

1. If PL is small, or PW is small, it is always Setosa.
2. If PL is small-medium or medium-large, and PW is small-medium or medium-large, it will be always recognized as Versicolor.
3. If PL is medium-large or large and PW is medium-large or large, then the class will be Virginica.

From the above three observations it is clear that the first observation classifies the Setosa correctly. However, for the classification of Versicolor and Virginica, there is overlap of input parameters in several sets like Large, Medium Large and Small Medium.

Therefore, it is clear there are no crisp rules which can classify the iris data correctly. We evolved fuzzy rules using grammatical evolution to classify the iris data.

<table>
<thead>
<tr>
<th></th>
<th>SL</th>
<th>SW</th>
<th>PL</th>
<th>PW</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>4.300</td>
<td>5.650</td>
<td>2.000</td>
<td>2.900</td>
</tr>
<tr>
<td>SM</td>
<td>4.750</td>
<td>6.550</td>
<td>2.300</td>
<td>3.500</td>
</tr>
<tr>
<td>ML</td>
<td>6.650</td>
<td>7.450</td>
<td>3.900</td>
<td>4.100</td>
</tr>
</tbody>
</table>
In order to view the correctness of the fuzzy rules evolved using grammatical evolution as shown in Table 1, we divided the vertical axis of each parameter into four uniformly distributed and overlapping fuzzy sets of small, small medium, medium large and large. Table 2 shows the range of fuzzy sets S, SM, ML and Large for each of the four attributes such as SL, SW, PL and PW.

Using parallel coordinates, we have divided the range of each input attribute such as SL, PL, PW and SW into fuzzy sets of S, SM, ML and Large as shown in Fig 6. All the eight rules are shown in Fig 6a-6h. Fig 6a represents rule 1 and was created using filters where SL is Large and PL is Large while the other two parameters, namely SW and PW are don’t care. The output category of this rule is Virginica. However, Fig. 6a shows three misclassifications giving the category as Versicolour, which indicates the strength of the rule is not 100%.

In fuzzy rules the strength of each rule varies between the interval 0 and 1. Therefore this only reflects that the strength of rule 1 is less than 1.

All the eight rules are taken together and applied to 150 data entries to see the overall classification success rate. Similarly, Fig 6b illustrates the correctness of the rule 2. The petal length belongs to the range of small medium and petal width ranges in the large scale. The rule 3 can be verified by applying filters on PL small, SW is medium-large SL is small and regardless of petal width as shown in Fig 6.c.

Fig. 6d verifies the correctness of rule 4 to predict Setosa. Fig. 6e represents the rule 5 and three input patterns converge to Virginica and other three patterns converge to Versicolour, which illustrates the rule strength of this rule, is 50%. On the contrary, Fig 6f illustrates the rule strength of rule 6 to be 100% as all the input patterns have been classified correctly to Virginica.

Furthermore, Fig. 6g shows the correctness of the rule 7. Rule 8 is represented in Fig. 6h. Here there are 8 patterns which are misclassified and only three patterns are correctly classified. This indicates the rule strength of this rule is 27%.

5. Analyzing the Rules in Parallel Coordinates

Our next objective is to verify the correctness of the fuzzy rules evolved by grammatical evolution. In this section we are visually verifying the accuracy of eight fuzzy rules as shown in Table 1. To verify the classification of each rule, we have transformed the input tuples to segmental parallel coordinate representation, which represent the fuzzy sets of the input parameters.

While the other two parameters, namely SW and PW are don’t care. The output category of this rule is Virginica. However, Fig. 6a shows three misclassifications giving the category as Versicolour, which indicates the strength of the rule is not 100%.

In fuzzy rules the strength of each rule varies between the interval 0 and 1. Therefore this only reflects that the strength of rule 1 is less than 1.

All the eight rules are taken together and applied to 150 data entries to see the overall classification success rate. Similarly, Fig 6b illustrates the correctness of the rule 2. The petal length belongs to the range of small medium and petal width ranges in the large scale. The rule 3 can be verified by applying filters on PL small, SW is medium-large SL is small and regardless of petal width as shown in Fig 6.c.

Fig. 6d verifies the correctness of rule 4 to predict Setosa. Fig. 6e represents the rule 5 and three input patterns converge to Virginica and other three patterns converge to Versicolour, which illustrates the rule strength of this rule, is 50%. On the contrary, Fig 6f illustrates the rule strength of rule 6 to be 100% as all the input patterns have been classified correctly to Virginica.

Furthermore, Fig. 6g shows the correctness of the rule 7. Rule 8 is represented in Fig. 6h. Here there are 8 patterns which are misclassified and only three patterns are correctly classified. This indicates the rule strength of this rule is 27%.
Fig. 6. Parallel coordinates of Fuzzy Classification Rule base on GE
6. Conclusion

In this paper, we presented the Parallel Coordinate Technique to visualize the fuzzy rules for classification of iris. This serves as a proof of concept for visualization of fuzzy classification system. By applying filters to the output, it is possible to generate some knowledge in the form of rules. This knowledge is utilized to classify the different species of Iris flower. Future work will include the application of this technique to achieve solutions for more complicated classification problems and investigating ways of obtaining higher classification rates.

7. References