Mining Interesting Regions using an Evolutionary Algorithm

J.L. Álvarez
D.I.E.S.I.A.
Universidad de Huelva
Huelva, Spain
alvarez@uhu.es

J. Mata
D.I.E.S.I.A.
Universidad de Huelva
Huelva, Spain
mata@uhu.es

J.C. Riquelme
D.L.S.I.
Universidad de Sevilla
Sevilla, Spain
riquelme@lsi.us.es

ABSTRACT

In this paper, we offer a new method to induce interesting knowledge from the relevant sets of data in databases for supervised learning. Thus, in this work, ELLIPSES is presented as a new method oriented to discover knowledge according to the expert's needs, by the detection of the most significant regions. The method essence is found in an evolutionary algorithm that finds these regions one after another. The expert decides which regions are significant and determines the stop criterion. The extracted knowledge is offered through two types of rules: Quantitative and Qualitative. The tool also offers a visualization of each rule by parallel coordinate systems. The ELLIPSES results are compared with C4.5 on UCI Repository datasets.

Keywords

Data Mining, Supervised Learning, Evolutionary Algorithm

1. INTRODUCTION

Nowadays, generally, the Knowledge Discovery in Databases (KDD) and, particularly, Data Mining (DM) have spurred a tremendous interest in the researchers community [1]. News algorithms and tools have been developed to Data Analysis (DA). Classification is an useful technique for discovering interesting rules in databases.

Classification systems are supervised learning methods that analyze a database or training set to build a classification model. The training set contains a feature collection, or object attributes whose class labels are known. The classification model is a set of rules for each class based on the data characteristics. Such rules are used to classify future objects according to the value of their attributes.

These methods are very useful and features have been justified with tools that have shown excellent results. But these techniques have some problems: they do not allow expert's intervention for learning process. So, classification systems normally generate a big number of rules whose interpretations are difficult. In these cases, the results are useless for an human-expert. Thus, it is necessary to include other techniques because the DA system main feature is to offer an easy interpretation of induced knowledge.

This paper presents ELLIPSES, a tool that permits to induce a set of classification rules in numerical attribute space [9][2]. These rules determine the most significant regions of the search space, they are a easy interpretation and the expert can also control the learning process, establishing when a region is interesting and the stop criterion.

Regions searching process is made by an evolutionary algorithm whose result interpretation is given by ELLIPSES through two rule models: quantitative and qualitative. It also offers a view of them using parallel coordinate systems so the relationship among attributes is shown by an image for each rule.

The rest of the paper is organized as follow. The mathematical preliminaries are presented in section 2. Then, section 3 describes ELLIPSES algorithm. And, in section 4 is shown the performance of our tool. This section offers the experimental results on Iris dataset and a comparison with C4.5 [13] on UCI Repository datasets [12]. The objective of this comparison is to offer a nexus between the classification systems and our tool, since our tool is not really a classification system. Finally, section 5 offers the conclusions about this method.

2. PRELIMINARIES

Our method uses conical regions to find the most significant rules. These regions contain the features of each class. This section offers the basic definitions of the models of rules used in our tool.

Definition 1. Let be an *hyperellipse* the natural extension of an ellipse in a d-dimensional space R^d.

Definition 2. Let be an *hyperellipsoide* the volume that is inside of an *hyperellipse*.

An hyperellipse (the wrapper) is equal to an ellipses or circumference in a two-dimensional space R^2. An hyperellipsoide (the wrapped volume) is equal to an ellipsoid or circle in a two-dimensional space R^2. Figure 1 offers a graphical representation of these concepts. Figure 1a) represents an ellipse of center (c_1, c_2), greater axis a_1 and smaller axis a_2 to two attributes x_1 and x_2 (two-dimensional space R^2)

SAC '2002 Madrid, Spain
Copyright 2002 ACM 1-58113-445-2/02/03 ...$5.00.

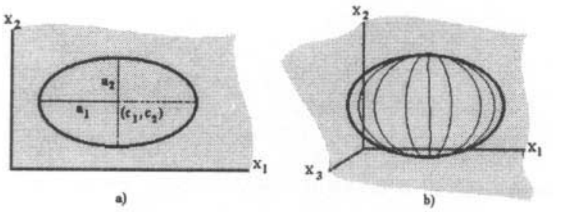

Figure 1: Graphical representation of an ellipse.

and figure 1b) shows an hyperellipse to three attributes x_1, x_2 and x_3 (three-dimensional space R^3).

$$\frac{(x_1-c_1)^2}{a_1^2}+\frac{(x_2-c_2)^2}{a_2^2}=1 \qquad (1)$$

$$\frac{(x_1-c_1)^2}{a_1^2}+\frac{(x_2-c_2)^2}{a_2^2}\leq 1 \qquad (2)$$

$$\frac{(x_1-c_1)^2}{a_1^2}+\frac{(x_2-c_2)^2}{a_2^2}+...+\frac{(x_d-c_d)^2}{a_d^2}\leq 1 \qquad (3)$$

The equation of the ellipse in R^2 is shown in 1. The equation of an ellipsoide is shown in 2. This equation is obtained changing = by $\leq$ in the equation of the associated ellipses. Generalizing, in R^d, the equation of an hyperellipsoide is shown in 3.

$$If\ x_1(c_1,a_1)\ and\ ...\ and\ x_d(c_d,a_d)\Rightarrow C_i \qquad (4)$$

$$h(x_i,a_i)=\begin{cases} Large\ if\ a_i>40\%A_x \\ MLarge\ if\ 25\%A_x<a_i\leq 40\%A_x \\ Medium\ if\ 15\%A_x<a_i\leq 25\%A_x \\ MShort\ if\ 5\%A_x<a_i\leq 15\%A_x \\ Short\ if\ a_i\leq 5\%A_x \end{cases} \qquad (5)$$

$$If\ x_1(c_1)\ width\ E_1\ and\ ...\ and\ x_d(c_d)\ width\ E_n\Rightarrow C_i \qquad (6)$$

The models of the rules (quantitative and qualitative) used in our tool are based on 1, 2 and 3. Thus, the quantitative model is obtained directly by the equation of the ellipse. This model is shown in 4 and it offers the central c_i value and the extent (width) a_i for each attribute, and the associated class C_i. The qualitative model uses five labels to specify the extent. For each attribute x_i, a E_j label is generated by $h(x_i,a_i)$ function, according to 5, where A_x is $x_{iM}-x_{im}$, x_{iM} is the maximum and x_{im} the minimum for x_i attribute. The qualitative model is shown in 6. The interpretation of these models of rule is very intuitive because the rule does not differ from the typical classification rules. Thus, let be $t:(y_1,y_2,...,y_n)$, if $y_i\in[x_i-a_i,x_i+a_i]\forall i$ then the item t is associated with the class C_i, according to 4. In the qualitative model, the label establishes the difference between y_i and x_i.

The method used to obtain the class C_i of an hyperellipsoide will be presented in the next sections, but this section offers the basic idea. Let be $t:(x_1,x_2,...,x_d,C_i)$ item, if i satisfies the equation 3 then the item is within the volume of the hyperellipsoide. Thus, the majority class within the hyperellipses is the associated class to it.

```
ELLIPSES Algorithm
    T ← Read Training set
    Repeat
        iter ← iter + 1
        $P_i$ ← Inicialice population on T
        Repeat
            Evaluate $P_i$ on T
            Select the best in $P_i$ to $P_{i+1}$
            Select 10% in $P_i$ to $P_{i+1}$
            Crossover $P_i$ individuals to $P_{i+1}$
            Mutate $P_{i+1}$
            $P_{i+1}$ is $P_i$
        Until number generations
        r ← Select the best of $P_i$
        if alpha(r)>ALPHA THEN add(R, r)
    Until (iter=ITER or beta(R)>BETA)
    Show R rules
    Visualization R by Parallel Coordinates
END.
```

Figure 2: ELLIPSES Algorithm

3. ELLIPSES ALGORITHM

The main objective of our tool is to induce the search space regions with a greater number of the items belonging to the same class and to permit the human-expert interaction in order to establish some criteria for the search process. The final result shows a reduced and easily interpretable set of rules. ELLIPSES is a DA tool based on Evolutionary Algorithm (EA) [5][6][11]. EAs are a heuristic search technique that has demonstrated to be robust for a variety of complex search space [4][14].

The technique maintains a population of individuals where each individual encodes a feasible solution to the problem. Iteratively, a new population is generated by replacing the previous population, according to Darwin's survival principle. So, each individual is evaluated to give its relative merit (fitness) as a solution. The new populations result from selection, crossover and mutation of previous populations. The evolutionary process is iterated by a predefined number of generations. The best individual of the evolutionary process is the solution of the algorithm

The EA has been used with excellent results [3][8][10]. In our method, a region is a conical surface. An EA is used to obtain the best regions. Figure 2 shows the ELLIPSES algorithm.

Iteratively, the EA finds the best hyperellipse r based on the number of positive and negative items in the hyperellipsoide. Let be $alpha(r)$ the percentage of the same class items in r, if $alpha(r)$ is greater than the predefined human-expert percentage $ALPHA$, then region r is considered. This process is repeated until reaching a predefined human-expert number of rules or predefined human-expert percentage $BETA$. Finally, the rules are shown according to 4 and 6 (quantitative and qualitative models), and they are shown by parallel coordinate systems.

3.1 Data structure of the individuals

An individual (a feasible solution) is a set $I=\{c_1,...,c_d,a_1,...,a_d\}$ where d is the number of attributes and $c_i,a_i\in\Re$ are the center and extent of the x_i attribute and they represent the equation of an hyperellipsoide according to 3.

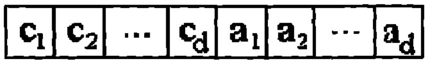

Figure 3: Representation of an individual.

Figure 3 shows a graphical representation of the individuals.

In practice, an individual represents a search space region. Each region will be associated to a class that will be deduced by the majority class of the data items in the hyperellipsoide.

3.2 Initial and next population

The process to generate the initial population consists in selecting, in a random way, each center c_i from the attributes range x_i ($[x_{im}, x_{iM}]$) and each extent a_i between 5% and 30% of the attributes range x_i.

The evolutionary process includes elitist: the best individual of every generation is replicated to the next one. Individuals are obtained through the copies of the previous population. These individuals are randomly and proportionally selected to their relative merit as a solution (fitness). The population remaining is formed through crossovers. Afterwards, mutation is applied depending on a probability.

3.3 Fitness function

The fitness (or merit as a solution) of an individual is obtained by training set item analysis. An item can be in or out of the hyperellipse. The out items are ignored. The different classes of the items in the hyperellipse are counted and the associated class to the individual is the majority class. Thus, the items with the same class are positive cases and the items with different classes are negative cases.

Furthermore, next iteration must direct the evolutionary process to other regions. Thus, the positive cases covered by discovered rules are considered covered cases. Finally, our method needs to obtain the greatest region. Thus, the amplitude of the hyperellipse is the hyperellipse volume divide by search spaces volume.

$$f(i) = Pos(i) - Neg(i) - Cover(i) * FC + Ampl(i) \quad (7)$$

Our algorithm maximizes the fitness function f for each individual i. The fitness function is given in 7, where $Pos(i)$ and $Neg(i)$ are the positive and negative cases in the hyperelipsoide that represent the individual i, $Cover(i)$ are the covered cases by previous hyperellipses, FC is the coverture factor and $Ampl(i)$ is the hyperellipse amplitude. Coverture factor (FC) is a value in the interval [0..1], and it offers the possibility of relaxing the covered cases, so, if FC is closed to 1, then the covered cases are considered negative cases, and if FC is closed to 0, then the covered cases are ignored.

3.4 Genetic operators

There are tree genetic operators: selection, crossover and mutation. To form a new population (the next generation), the individuals are selected according to their fitness by the selection operator. Many selection procedures are currently in use, our algorithm uses roulette wheel procedure, where individuals are selected with a proportional probability to their relative fitness. This ensures that an individual is chosen in a expected number of times approximately proportional to its relative performance in the population. Thus, high-fitness (good) individuals stand a better chance of selecting, while low-fitness individuals are more likely to disappear.

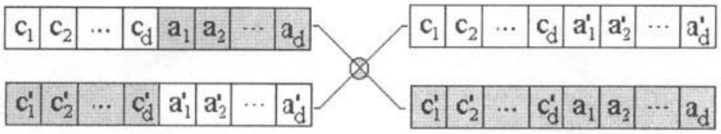
c_1 c_2 ... c_d a_1 a_2 ... a_d
c'_1 c'_2 ... c'_d a'_1 a'_2 ... a'_d
c_1 c_2 ... c_d a'_1 a'_2 ... a'_d
c'_1 c'_2 ... c'_d a_1 a_2 ... a_d

Figure 4: The middle point crossover operator.

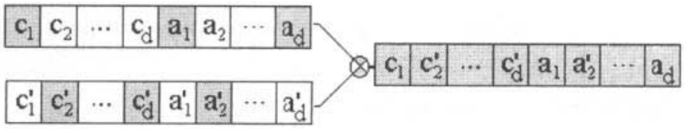
c_1 c_2 ... c_d a_1 a_2 ... a_d
c'_1 c'_2 ... c'_d a'_1 a'_2 ... a'_d
c_1 c'_2 ... c'_d a_1 a'_2 ... a_d

Figure 5: The uniform crossover operator.

Selection cannot introduce any new individuals into the population. These individuals are generated through crossover and mutation operators. Crossover operator is performed by selecting two individuals called parents, and generating new individuals called offspring. In our algorithm, the crossover operator has two components: the middle point crossover and the uniform crossover. They are performed with a probability p_{cross} that chooses between the middle point crossover and the uniform crossover. The middle point crossover randomly splits the individuals in two parts. Then the fragments are exchanged generating two new individuals. Figure 4 graphically shows this process. The uniform crossover decides , independently for each coefficient of an individual, whether it contribute or not to the new individual. An example of this procedure is shown in figure 5.

$$v_{ij} = v_{ij} \pm Quant * PerMut * v_{ij} \quad (8)$$

Finally, the mutation operator is introduced to prevent premature convergence to local optimum by randomly sampling new points in the search space. Three variants are implemented: center mutation, amplitude mutation and extreme mutation. Mutation is performed with probability p_{mut} on an individual. When an individual must be mutated, a probability chooses between the different operators. The center and amplitude mutation operators alter the center $(c_1, ..., c_d)$ and the extent $(a_1, ..., a_d)$ of the hyperellipse, respectively, according to 8, where v_{ij} is the factor to alter, $Quant$ and $PerMut$ take their values from [0..1], $Quant$ is the random quantity that v_{ij} is altered and $PerMut$ is the percentage of mutation that determines how the mutation influence on v_{ij}. The extreme mutation operator alters both center (c_i) and extent (a_i) of an attribute (x_i). Thus, the mutation let the middle value of $x_{iM} - x_{im}$ to c_i and let $\frac{x_{iM}-x_{im}}{2}$ to a_i. The objective of this operator is to cover the attribute.

3.5 Parallel coordinate systems

Although our tool offers two models of rules and the qualitative model is easily interpreted, sometimes it is necessary to provide the information using another philosophy. Thus, a visualization of the relationships among the attributes offers a good support to the expert. The visualization technique used in our algorithm is shown in this section. This

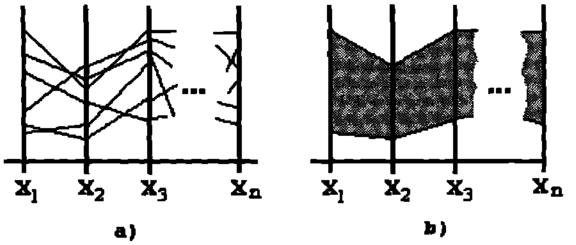

Figure 6: Parallel Coordinate Systems.

technique offers the relationships among attributes by parallel coordinates [7].

A parallel coordinate system is composed by a set of parallel axes separated by a fixed distance. Each axis corresponds with an attribute and they are escalated on the range of the attribute. Thus, d axes are necessary to represent d attributes. In this system, a line represents each data item. This line intersects with each axis on the value of the item for that attribute. Figure 6a) shows the traditional parallel coordinate system.

In our method, each region is represented on a parallel coordinate system. But, all data items in a region are not represented on parallel coordinate system. Thus, only the minimal value and the maximal value, for each attribute, are represented on each axis and these values are joined by filled polygonal. Figure 6b) offers an example of this method. The internal lines are eliminated. The objective of this variant is to offer a clearer and compact vision of the relationships between the attributes.

4. RESULTS

In order to evaluate the performance of our tool this section offers the results on UCI Repository datasets [12]. Thus, it shows the obtained rules and their visualization on parallel coordinate systems on the tradicinal Iris dataset in section 4.1. Furthermore, section 4.2 offers a comparison between ELLISPES and C4.5.

4.1 Iris dataset

To illustrate the results induced by ELLIPSES, this section offers the results that has been discovered on Iris dataset.

$$\begin{array}{l} If\ pw(0.3, 0.67) \Rightarrow Set.(50/0/0)(33\%) \\ If\ pw(2.5, 0.73) \Rightarrow Vir.(45/1/0)(30\%) \\ If\ pl(3.9, 0.94) \Rightarrow Ver.(46/3/0)(31\%) \end{array} \tag{9}$$

$$\begin{array}{l} If\ pw(0.3)\ width\ MLARGE \Rightarrow Set.(50/0/0) \\ If\ pw(2.5)\ width\ MLARGE \Rightarrow Vir.(45/1/0) \\ If\ pl(3.9)\ width\ MEDIUM \Rightarrow Ver.(46/3/0) \end{array} \tag{10}$$

The quantitative model of the rules is shown in 9. The interpretation is very intuitive although this is the quantitative model. So, for example, the first rule shows that if the pw (petal width) attribute is round of 0.3 with an extent of ± 0.67 then the obtained class is Iris-Setosa. The qualitative model is shown in 9. This model uses a label to represent the amplitude. This label offers qualitative information of the amplitude of the rule on an attribute respect to the range of the attribute, according to 5. The rules show the number of positive, negative and covered cases and the percentage of positive cases.

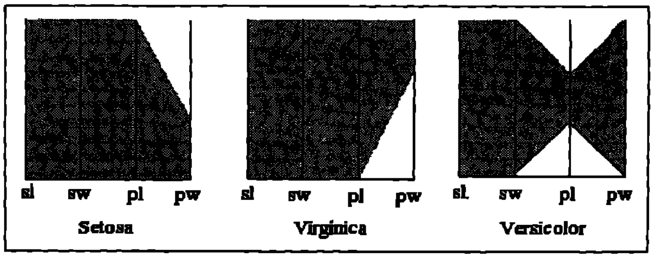

Figure 7: Iris Dataset Visualization.

The visualization of this rules by parallel coordinate systems is shown in figure 7. This representation offers a graphical description of the previous rules. For example, it shows the following: if pw takes short values the class is Iris-setosa, if pw takes high values then the class is Iris-Virginica and, in other case, if pl takes middle values then the class is Iris-Versicolor. This visualization offers a very intuitive and easy interpretation of the rules.

4.2 ELLIPSES vs C4.5: A comparison

This section offers a comparison of the results of ELLIPSES versus C4.5. Though ELLIPSES is not a classification system, a method is presented in order to evaluate our tool. This method compares the results obtained by ELLIPSES with the results obtained by C4.5 on six UCI Repository datasets. The features of these datasets are shown in table 1.

For this, it offers a comparison based on the number of rules obtained by ELLIPSES. Thus, table 2 shows the percentage of positive cases of each class (column *%cls*), the percentage of positive cases on the total (column *%ttal*) and the percentage of negative cases or error rate (column *%er*). As C4.5 is a classification system, it finds more rules than our tool thus the most meaningful rules (column *r*) are only used in the comparison. The rules that more items collect are the most meaningful rules.

To clarify the content of table 2, we offer an explanation of the results on PIMA dataset. This dataset has 768 items, 8 attributes and two class denoted with 0 and 1, as table 1 shows. ELLIPSES induces two rules for the class 0. These rules cover 52.2% of the items of the 0 class, this is 33.9% on all items and the percentage of error is 2.8%.

C4.5 induces 15 rules for class 0. In this comparison the two rules that cover more items are considered. The two most meaningful rules cover 54.8% of the items of the 0 class, 35.6% of all items and the percentage of error is 2.9%. In a same way, ELLIPSES induces a rule for class 1 that covers 20.5% of the class, 7.2% of all items and 0.6% error. C4.5 induce 7 rules where the most meaningful rule covers 29.8% of the class, 10.4% of all items and 1.5% error.

The previous results show that the accuracy of the classification in both methods is very similar, although the most significant rules are only used. Furthermore, the rate of error is lightly inferior in ELLIPSES. These results determine that ELLIPSES is a good tool to obtain interesting rules (regions). Furthermore, ELLIPSES has other advantage: "human expert's interaction". Thus, human experts can determine the number, the support and the confidence of the rules. That is to say, they determine the importance of the regions.

Table 1: UCI Repository Datasets

Datasets	#Items	#Att.	#Class	Class
BCW	699	10	2	2,4
BUPA	345	6	2	1,2
GLASS	214	9	7	1,2,3,5,6,7
HAYES	132	5	3	1,2,3
IRIS	150	4	3	s,v,i
PIMA	768	8	2	0,1

Table 2: ELLIPSES vs. C4.5: A Comparison

			ELLIPSES			C4.5		
	c	r	%cls	%ttal	%er	%cls	%ttal	%er
BCW	2	1	90.1	59.1	0.2	81.0	53.0	0.1
	4	1	71.4	24.6	0.4	67.6	23.3	0.3
BUPA	1	3	32.4	13.6	1.1	36.5	15.3	1.4
	2	5	40.0	23.1	0.2	74.5	43.1	11.8
GLASS	1	3	81.4	26.6	0.4	78.5	25.7	0.4
	2	3	64.4	22.8	2.8	64.4	22.8	3.2
	3	2	64.7	5.1	0.4	47.0	3.7	0.0
	5	1	76.9	4.6	0.4	92.3	5.6	0.4
	6	1	66.6	2.8	0.0	100.0	4.0	0.0
	7	1	82.7	11.2	0.0	93.1	12.6	0.4
HAYES	1	3	68.6	26.5	0.0	64.7	25.0	0.0
	2	3	64.7	25.0	1.5	62.7	24.2	0.7
	3	3	100.0	22.7	0.0	100.0	22.7	0.0
IRIS	s	1	100.0	33.3	0.0	100.0	33.3	0.0
	v	1	94.0	31.3	0.0	94.0	31.3	0.1
	i	1	88.0	29.3	0.0	90.0	30.0	0.1
PIMA	0	2	52.2	33.9	2.8	54.8	35.6	2.9
	1	1	20.5	7.2	0.6	29.8	10.4	1.5

Table 3: Parameters and default values

Parameter	Default value
ALPHA	10.0%
BETA	90.0%
ITER	10
num Generations	200
num Individuals	200
% selected individuals	10%
Probability of crossover	50%
Probability of mutations	33%
Percentage of mutation	70%

As disadvantages, our tool has the handicap of the evolutionary computation: high computational cost. However, in this case the results show that in relatively few generations, the found regions are sufficiently valid. It is necessary to know that the final purpose of our tool is not a classification system that optimizes the error rate, since the purpose is to find qualitatively interesting regions.

Table 3 shows the fundamental parameters and their defaults values used to induce the previous results.

5. CONCLUSION

In this paper, we present a new supervised learning tool into DM field. The main objective is to induce a set of rules (knowledge) about qualitative interesting regions on a database. These rules are easier to interpret for a human expert because they are shown via three formats: quantitative, qualitative and parallel coordinate systems. Furthermore, this tool permits humans experts interaction by the definition of parameters in the learning process.

Analyzing the previous section, it can be deduced that ELLIPSES is not a classification system, since their main objective is not to optimize the error rate. Thus, the obtained results are not the same that the results of a classification system, as C4.5. But, without any doubt, analyzing also the results in table 2, we can conclude that ELLIPSES acts as a classification system when the objective is to find the most interesting regions, rather the rules that determine regions with an interesting volume of items.

Summarizing, our tool offers the human interaction in the learning process, two models of rules: quantitative and qualitative, and a visualization by parallel coordinate systems, so we can conclude that it is an excellent tool in data mining field.

6. ACKNOWLEDGMENTS

This work has been supported by Spanish Research Agency CICYT under grant TIC2001-1143-C03-02

7. REFERENCES

[1] M.-S. Chen, J. Han, and P. S. Yu. Data mining: an overview from a database perspective. *IEEE Tr. On Knowledge And Data Engineering*, 8(6):866–883, 1996.

[2] T. V. de Merckt. Decision trees in numerical attribute spaces. In *IJCAI*, pages 1016–1021, 1993.

[3] K. A. DeJong, W. M. Spears, and D. F. Gordon. Using genetic algorithms for concept learning. *Machine Learning*, 13(2/3):161–188, 1993.

[4] L. Eshelman and J. Schaffer. Real-coded genetic algorithms and interval schemata, 1993.

[5] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

[6] J. H. Holland. *Adaptation in natural artificial systems.* University of Michigan Press, Ann Arbor, 1975.

[7] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–92, 1985.

[8] C. Z. Janikow. A knowledge-intensive genetic algorithm for supervised learning. *Machine Learning*, 13:189–228, 1993.

[9] M. Kasif and S. Merckt. Nfdt: A system that learns flexible concepts based on decision trees for numerical attributes, 1992.

[10] J. Koza. Concept formation and d.t. induction using the genetic programming paradigm, 1991.

[11] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs, Third Edition.* Springer, Berlin, 1999.

[12] P. Murphy and D. Aha. UCI Repository of machine learning databases, 1992.

[13] J. Quinlan. C4.5: Programs for machine learning, 1993.

[14] A. Wright. Genetic algorithms for real parameter optimizations. In *Morgan Kaufmann Pub.*, pages 205–218, 1991.