Decreasing Occlusion and Increasing Explanation in Interactive Visual Knowledge Discovery

Boris Kovalerchuk(✉) and Abdulrahman Gharawi

Central Washington University,
400 E. University Way, Ellensburg, WA 98926, USA
{BorisK, Abdulrahman.Gharawi}@cwu.edu

Abstract. Explanation and occlusion are the major problems for interactive visual knowledge discovery, machine learning and data mining in multidimensional data. This paper proposes a hybrid method that combines the visual and analytical means to deal with these problems. This method, denoted as FSP, uses visualization of n-D data in 2-D, in a set of Shifted Paired Coordinates (SPC). SPCs for n-D data consist of n/2 pairs of Cartesian coordinates, which are shifted relative to each other to avoid their overlap. Each n-D point is represented as a directed graph in SPC. It is shown that the FSP method simplifies the pattern discovery in n-D data, providing the explainable rules in a visual form with a significant decrease of the cognitive load for analysis of n-D data. The computational experiments on real data has shown its efficiency on both training and validation data.

Keywords: Visual knowledge discovery · Visual analytics
Visual data mining · Occlusion · Interactive visualization
Shifted Paired Coordinates

1 Introduction

1.1 Background and Problem

For a long time, explanation and occlusion have been among the major problems for interactive visual knowledge discovery, data mining, and machine learning in multidimensional data. This paper proposes a hybrid method, which combines the visual and analytical means to deal with these problems. The proposed method, denoted as FSP, uses visualization of n-D data in 2-D in a type of General Line Coordinates (GLC) [1, 2] known as Shifted Paired Coordinates (SPC). A set of Shifted Paired Coordinates for n-D data consists of n/2 pairs of Cartesian coordinates that are shifted relative to each other without overlap. Each n-D point A is represented as a directed graph A^* in SPC, where each node of A^* is a 2-D projection of A in a respective pair of the Cartesian coordinates.

The proposed FSP method significantly decreases the cognitive load for the analysis of n-D data and simplifies the discovery of *explainable patterns* in the n-D data. At the upper level, the steps of the FSP are: (1) *Filtering* out less efficient visualizations

S. Yamamoto and H. Mori (Eds.): HIMI 2018, LNCS 10904, pp. 505–526, 2018.
https://doi.org/10.1007/978-3-319-92043-6_42

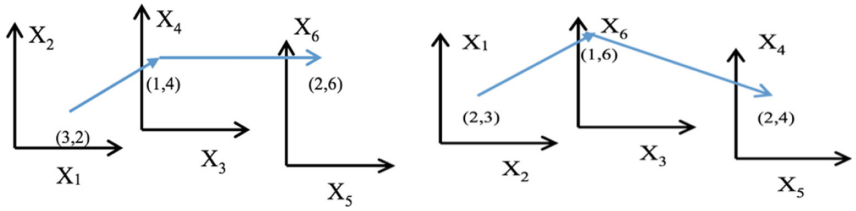
from multiple SPC visualizations, (2) *Searching* for sequences of paired coordinates that are more efficient, and (3) *Presenting* the SPC visualizations only with better sequences to the analyst. FSP includes the randomized search for pairs of coordinates and explainable "rectangular" classification rules with maximized accuracy on training/validation data.

The computational experiments with the 9-D Wisconsin Breast Cancer data, the 33-D Ionosphere data, and the 8-D Abalone data, from UCI Machine Learning repository, show the efficiency of the FSP method, on the training and validation data. The visualization process in SPC is reversible, i.e., all n-D information is visualized and can be restored from visualization for each n-D case. This hybrid visual analytics method allows classifying data in a way that can be communicated to the domain experts such as medical doctors in the explainable/understandable and visual form.

1.2 Shifted Paired Coordinates: Challenge and Opportunity to Better Visualization

The **Shifted Paired Coordinates** (**SPC**) visualization of the n-D data requires the splitting of n coordinates $X_1 - X_n$ into pairs producing the $n/2$ non-overlapping pairs (X_i, X_j), such as $(X_1, X_2), (X_3, X_4), (X_5, X_6), \ldots, (X_{n-1}, X_n)$ [1, 2]. In SPC, a pair (X_i, X_j) is represented as a separate orthogonal Cartesian Coordinates (X, Y), where X_i is X and X_j is Y.

In SPC visualization design each coordinate pair (X_i, X_j) is *shifted* relative to other pairs to avoid their overlap. This creates $n/2$ scatter plots. Next in SPC, for each n-D point $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the point (x_1, x_2) in (X_1, X_2) is connected to the point (x_3, x_4) in (X_3, X_4) and so on until point (x_{n-2}, x_{n-1}) in (X_{n-2}, X_{n-1}) is connected to the point (x_{n-1}, x_n) in (X_{n-1}, X_n) to form a directed graph $\mathbf{x}^*$. Figure 1 shows the same data visualized in SPC in two different ways due to different pairing of coordinates.



Point **a** in (X_1,X_2), (X_3,X_4), (X_5,X_6) as a sequence of pairs (3,2), (1,4) and (2,6).

Point **a** in (X_2,X_1), (X_3,X_6), (X_5,X_4) as a sequence of pairs (2,3), (1,6) and (2,4).

Fig. 1. 6-D point **a** = (3, 2, 1, 4, 2, 6) in Shifted Paired Coordinates.

In general, there are multiple combinatorial ways to form the pairs of coordinates for the SPCs, and to sequence the pairs. The SPC visualization graphs $\mathbf{x}_k^*$ of each given n-D point **x** differ for different sequences S_k of pairs of coordinates. Figure 1a illustrates

it for a 6-D point **a** = (3, 2, 1, 4, 2, 6) visualized in pairs (X_1, X_2), (X_3, X_4), (X_5, X_6), and Fig. 1b shows this point visualized in pairs (X_2, X_1), (X_3, X_6), (X_5, X_4).

The SPC allows visualizing each individual n-D point losslessly, but together graphs of multiple n-D points occlude each other. See Fig. 2. This creates a difficulty for discovering patterns to classify cases of opposing classes in SPC. In Fig. 2, some of the areas are visibly dominated by the cases of the specific color. However, it is not sufficient to build discrimination rules to classify cases visually. It required an addition analytical process. Such process is proposed in this paper. SPC visualizations with some sequences S_k can reveal classification patterns of n-D data better than with using other sequences. The *dependence* of the visualization on the different pairing of coordinates creates a *challenge* and an *opportunity* to find the better pairs and their sequences.

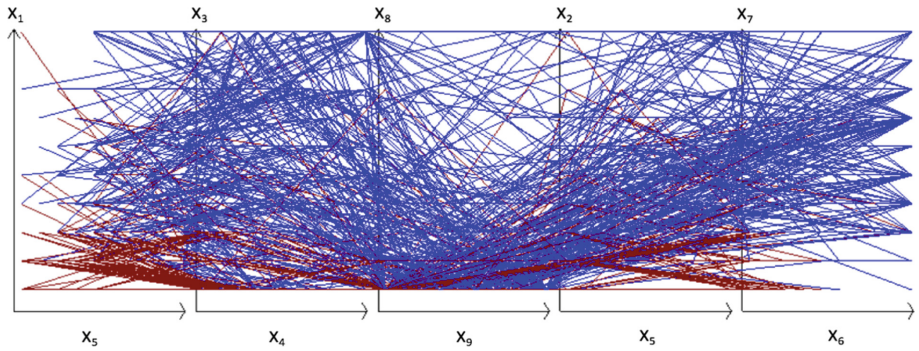


Fig. 2. A set of 688 Wisconsin Breast Cancer (WBC) data visualized in SPC as 2-D graphs of 10-D points with benign cases in red and malignant cases in blue. (Color figure online)

The challenge is that it is impractical to conduct the interactive search of the efficient sequences of pairs of coordinates for the large number of sequences. The total number of pairs of n coordinates is the number of combinations $C(n,2) = n!/((n-2)! \cdot 2!)$, e.g., 45 for n = 10. Next, there are multiple different sequences of the same set of n/2 pairs, e.g., $(X_1, X_2), (X_3, X_4), (X_5, X_6), \ldots, (X_{n-1}, X_n)$ and $(X_5, X_6), (X_3, X_4), (X_1, X_2), \ldots, (X_{n-1}, X_n)$. The number of these sequences (orders) for the given n/2 pairs is $(n/2)!$, e.g., (10/2)! = 120 for n = 10. Thus, the total number of sequences of n/2 pairs of n coordinates is $(n/2)! \cdot C(n,2) = (n/2)! \cdot n!/((n-2)! \cdot 2!)$ and for n = 10 it is 45·120 = 5400 assuming that any of C(n,2) pairs can be used to form sequences. The analyst cannot observe all of them, to find the best one for visual separation of classes. The FSP algorithm resolves this issue by the automatic search for the best sequences, and presenting only a few of the best visualizations to the analyst.

2 Algorithms

2.1 FSP Algorithm

The FSP algorithm:

(a) *Filters* out less efficient rules and visualizations for supervised classification learning,
(b) *Searches* for sequences of pairs of coordinates and respective rules that are more efficient for supervised classification learning, and
(c) *Presents* the SPC visualizations only with better sequences to the analyst.

The main characteristic of the FSP is avoiding the interactive exploration of the exponential number of alternative sequences with the following *major steps*:

Step 1: *Random generation* of sequences of pairs of coordinates *S*;
Step 2: *SPC representation* of n-D data in sequences *S* from Step 1;
Step 3: *Machine learning* process for learning "rectangular" classification rules with high accuracy, precision and recall in sequences *S* from Step 2;
Step 4: *Full* automated visualization process: SPC representation of best n-D rules in the best *full sequences S* of pairs of coordinates *S* discovered in Step 3;
Step 5: *Simplified* automated visualization process: *SPC representation* of best n-D rules in the best *subsequences S* of pairs of coordinates *S* discovered in Step 3;
Step 6: *Interactive* visualization process: an analyst interactively controls and manages produced SPC visualizations.

The approach of this algorithm is in line with [4]. It produces efficient classification rules and 3-D visualization of n-D data. In that paper another visualization method called Collocated Tripled Coordinates from the class of General Line Coordinates (GLC) is combined with Machine Learning for predicting the investment strategy. The ideas of steps 1 and 2 already have been explained above. The step 3 uses rules and learning criteria presented below.

Rules and Learning Criteria. The filtering works on a set of rules such as the rules (1)–(8), listed below. Each rule is defined on an n-D point $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to be classified into some classes:

$$\text{If } (x_i, x_j) \in R_1 \text{ then } \mathbf{x} \in \text{class } 1, \tag{1}$$

$$\text{If } (x_i, x_j) \in R_1 \,\&\, (x_k, x_m) \in R_2 \,\&\, (x_s, x_t) \in R_3 \text{ then } \mathbf{x} \in \text{class } 1 \tag{2}$$

$$\text{If } ((x_i, x_j) \in R_1 \vee (x_k, x_m) \in R_2) \,\&\, (x_s, x_t) \in R_3 \text{ then } \mathbf{x} \in \text{class } 1 \tag{3}$$

$$\text{If } ((x_i, x_j) \in R_1 \vee (x_k, x_m) \in R_2) \,\&\, (x_s, x_t) \notin R_3 \text{ then } \mathbf{x} \in \text{class } 1 \tag{4}$$

$$\text{If } ((x_i, x_j) \in R_1 \vee (x_k, x_m) \in R_2) \,\&\, (x_s, x_t) \notin R_3 \text{ then } \mathbf{x} \in \text{class } 1, \text{ else } \mathbf{x} \in \text{class } 2 \tag{5}$$

$$\text{If } (x_i, x_j) \in R_1 \,\&\, (x_k, x_m) \in R_2 \,\&\, (x_s, x_t) \notin R_3 \text{ then } \mathbf{x} \in \text{class 1} \tag{6}$$

$$\text{If } (x_i, x_j) \in R_1 \,\&\, (x_i, x_j) \notin R_2 \,\&\, (x_i, x_j) \notin R_3 \text{ then } \mathbf{x} \in \text{class 1} \tag{7}$$

$$\text{If } ((x_i, x_j) \in R_1 \vee (x_k, x_m) \notin R_2 \vee (x_s, x_t) \notin R_3 \text{ then}) \, \mathbf{x} \in \text{class 1} \tag{8}$$

where R_1, R_2 and R_3 are specific *rectangles*, in respective pairs of Cartesian coordinates in a given sequence S of pairs of coordinates S, e.g., R_1 can be in (X_1, X_2).

The filtering follows the common Data Mining/Machine Learning strategy of learning the rules on the training data, and testing on the validation data. The *quality of learning of the classification* and *the expected visualization* for the rules (1)–(4), (6)–(8) is measured by the **precision** and **recall** of the classification of the training and validation data, where the precision Pr is the fraction of the of cases, predicted correctly by the rule, over all the predicted cases by the rule:

$$Pr = |\{\text{cases predicted correctly by the rule}\}|/|\{\text{all predicted cases by the rule}\}|.$$

The precision Pr for the basic rule (1): If $(x_i, x_j) \in R_1$ then $\mathbf{x} \in$ class 1, is calculated as follows:

$$Pr = \frac{n1(R_1)}{n1(R_1) + n2(R_1)} \tag{9}$$

where n1(R_1) is the number of points of class 1 in R_1 (i.e., the number of correctly classified cases), and n2(R_1) is the number of points from the class 2 in R_1 (i.e., the number of misclassified cases). More generally, for any rule($\mathbf{x}$) such that

$$\text{If rule}(\mathbf{x}) \text{ then } \mathbf{x} \in \text{class 1} \tag{10}$$

the precision is

$$Pr = \frac{n1(rule)}{n1(rule) + n2(rule)} \tag{11}$$

where $n1$(rule) is the number of points of class 1 that satisfy the if part of the rule and $n2$(rule) is the number of points of class 2 that satisfy the if the part of the rule too.

The formula (11) is applicable to all the rules (1)–(8). For example, the precision Pr for the rule (4) is calculated as follows:

$$Pr = \frac{n1(R_1) + n1(R_2) - n1(R_1 \& R_3) - n1(R_2 \& R_3)}{n1(R_1) + n1(R_2) - n1(R_1 \& R_3) - n1(R_2 \& R_3) + n2(R_1) + n2(R_2) - n2(R_1 \& R_3) - n2(R_2 \& R_3)} \tag{12}$$

where

$n1(R_1)$ and $n1(R_2)$ are the number of points of class 1 in R_1, R_2, respectively,

$n1(R_1 \& R_3)$ is the number of graphs $\mathbf{x}^*$ of the n-D points of class 1 that have 2-D points in both R_1 and R_3,

$n1(R_2\&R_3)$ is the number of graphs $\mathbf{x}^*$ of the n-D points of class 1 that have 2-D points in both R_2 and R_3,

$n2(R_1\&R_3)$ is the number of graphs $\mathbf{x}^*$ of the n-D points of class 2 that have 2-D points in both R_1 and R_3,

$nb(R_2\&R_3)$ is the number of graphs $\mathbf{x}^*$ of the n-D points of class 2 that have 2-D points in both R_2 and R_3.

Here

$n1(R_1) + n1(R_2) - n1(R_1\&R_3) - n1(R_2\&R_3)$ is the number of correctly classified n-D points by rule (4) and

$n2(R_1) + n2(R_2) - n2(R_1\&R_3) - n2(R_2\&R_3)$ is the number of misclassified n-D points by this rule.

We use the precision for rules (1)–(4) and (6)–(8) instead of the accuracy, because these rules predict only one class. All cases, which do not satisfy the condition of these rules, are not classified (*refused* to be classified). The computing accuracy would require predictions of the class for all cases as it is the case for rules (5). Therefore, rules in set of rules (5) we use the accuracy for filtering. Note that a high precision rule from sets of rules (1)–(4), (6)–(8) may covers only a few cases, but the precision value does not show it's low coverage. Therefore, we also use the **recall**

$$RC = |\{\text{cases predicted correctly by the rule}\}|/|\{\text{all cases}\}|$$

that is a fraction of cases correctly predicted by the rule to all cases to be predicted.

The **Random generation** in Step 1 consists of the two substeps:

(RS1) Randomly generate a set of pairs of coordinates from coordinates $X_1 - X_n$,
(RS2) Randomly generate sequence S_k for a set of pairs from (RS1).

The **Machine Learning process** in the Step 3 consists of the following substeps:

(ML1) Search for rectangles in each (X_i, X_j) that maximize precision or accuracy of a rule from (1)–(8) on training data for given S_k,
(ML2) Evaluate this rule on validation data,
(ML3) Repeat (RS1)–(RS2) and (ML1)–(ML2) in attempt to reach desired precision/accuracy and recall.
(ML3) Combine promising rules to get a stronger rule in precision, accuracy and recall.

The **Automated Visualization process** in Steps 4–5 consists of the following steps:

(AV1) Visualize in SPC most accurate classification results. This includes visualization of only best results.
(AV2) Remove data that are covered by best results in (IV1),
(AV3) Repeat (RS1–RS2) and (ML1–ML3) for remaining data in search for the best classification results.

The **Interactive Visualization process** in the Step 6 is as follows:

(IV1) Substitute the automatic search in (ML1) by the interactive search, where the analysts select the rectangles in SPC visualization using the GUI.
(IV2) The automatic system supports this interactive process by computing accuracies of rules based on selected rectangles and removing data covered by best results found before the next interactive selection of new rectangles starts.

Figure 3 shows the overall design of the FSP process.

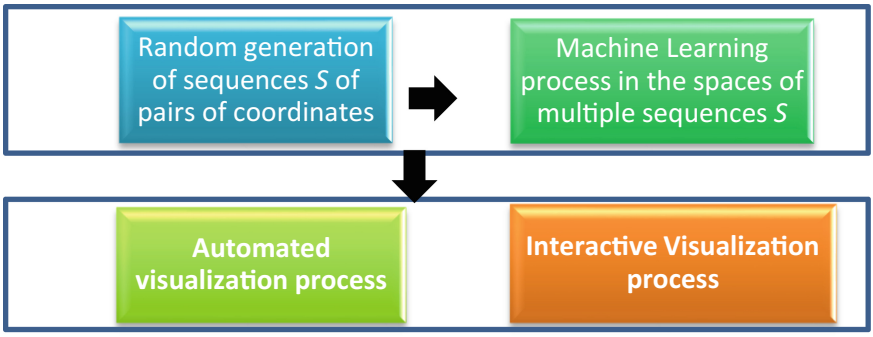


Fig. 3. Overall design of the FSP process.

2.2 Frequency Visualization Algorithm

One of the challenges of the SPC visualization for the larger data sets is a form of occlusion, caused by a limited resolution of the data on the screen. It leads to overlap of similar data including complete colocation of some lines. In addition, identical cases will collocate on any visualization at any resolution. Therefore, the task is enhancing the SPC visualization to show frequency of the lines.

The algorithm for this, denoted as the **FRE algorithm**, has the following steps:

(F1) For each consecutive pairs of coordinates (X_i,X_j), (X_k, X_m) form sets of edges $\{E_q\}$ that are collocated or nearly collocated under some threshold T,
(F2) Count the number of edges $C_q(T)$ in each of these sets,
(F3) Draw each such set of edges E_q with the adjustable width w:
  (a) *Equally* proportional to the number of edges in E_q for each node,
  (b) *Unequally* proportional to the number of edges in E_q for each node *without adjusting* to the data density focused in the given node of graph.,
  (c) *Unequally* proportional to the number of edges in E_q for each node *with adjusting t*o the data density in the given node of graph.

Figure 4 illustrates the differences between the versions (a)–(c) of the algorithm FRE for the red graphs in the SPCs. The version (a) is "neutral", the analyst can use it when no specific node it set up to explore, Versions (b) and (c) are for exploring

specific nodes of interest (e.g. nodes of the discovered classification rule), because the large width of edge in (a) may occlude that specific node.

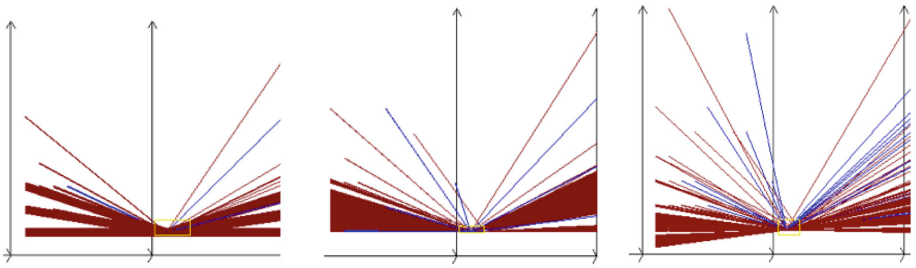


Fig. 4. Alternative frequency visualizations.

3 Experimental Case Studies

3.1 Experimental Case Study 1

The computational experiments with the 9-D Wisconsin Breast Cancer (WBC) data, from the UCI Machine Learning repository [3] presented below, show the *efficiency* of the FSP algorithm. To get the even number of coordinates and 5 pairs of coordinates, the coordinate X_5 was duplicated in X_{10} getting total 10 coordinates.

The discovered patterns were found by the search in the set of rules (1)–(8). In particular, on WBC data, the FSP algorithm found an efficient sequence of the pairs of the coordinates. This sequence of pairs is (X_5, X_1), (X_4, X_3), (X_9, X_8), (X_5, X_2), (X_6, X_7). Here X_5 is used in two pairs (X_5, X_1) and (X_5, X_2). The SPC visualization with this sequence reveals classification pattern with *precision over 90% i*n all 11 random 70%:30% splits that are presented in Tables 1 and 2. The best precision on the training data is 99.3%, which is accompanied by the high precisions on the validation data (98.21%), in one of the 70%:30% splits of the given data into the training and the validation data.

The discovered rules in Tables 1 and 2 belong to the set of rules (7). The first rule in Table 1 that we denote as **WBC Rule 1** is:

$$\text{If } (x_9, x_8) \in R_1 \,\&\, (x_6, x_7) \notin R_2 \,\&\, (x_6, x_7) \notin R_3 \text{ then } \mathbf{x} \in \text{class } 1(\text{Red}, \text{Benign}) \quad (13)$$

where R_1,R_2,and R_3 are specific rectangles, in respective pairs of Cartesian coordinates (X_8, X_9) and (X_6, X_7). Table 3 shows the parameters of R_1,R_2 and R_3 in the normalized coordinates.

This rule, with a random 70%:30% data split into the training and the validation data, has the precision of 94.6% on the training data, and 96.82% on the validation data. Figures 5 and 6 show its rectangles R_1, R_2 and R_3 drawn in the SPC as magenta boxes. The difference between these figures is that Fig. 5 shows all graphs that go through rectangle R_1 (have node (x_9, x_8) in R_1), but Fig. 6 shows only graphs that in

Table 1. Number of cases that satisfy the if part of Rule 1 in 11 random 70%:30% splits of data.

70%:30% random data splits	Number of cases that satisfy if part of the Rule 1					
	Red class			Blue class		
	Training	Testing	Total	Training	Testing	Total
1	303	122	425	17	4	21
2	300	105	405	10	4	14
3	290	132	422	20	4	24
4	291	110	401	2	2	4
5	253	123	376	2	1	3
6	297	116	413	13	3	16
7	301	121	422	12	2	14
8	299	122	421	12	4	16
9	282	127	409	10	4	14
10	307	116	423	27	7	34
11	282	117	399	27	9	36
Average	291	119	411	14	4	18

Table 2. Precision, recall and coverage of Rule 1 in 11 random 70%:30% splits of data.

70%:30% random data splits	Rule precision		Rule recall (correct) coverage			Rule total coverage of cases, %
	Training, %	Testing, %	Training, %	Testing, %	Total, %	
1	94.06	96.82	44.04	17.73	61.77	64.83
2	96.77	96.33	43.6	15.26	58.86	60.9
3	93.5	97.05	42.15	19.19	61.34	64.83
4	99.3	98.21	42.3	15.99	58.29	58.87
5	98.41	99.19	36.77	17.88	54.65	55.09
6	95.8	97.47	43.17	16.86	60.03	62.35
7	96.16	98.37	43.75	17.59	61.34	63.37
8	96.14	96.82	43.46	17.73	61.19	63.52
9	96.57	96.94	40.99	18.46	59.45	61.48
10	91.91	94.3	44.62	16.86	61.48	66.42
11	91.26	92.85	40.99	17.01	58	63.23
Average	95.44	96.76	42.3	17.3	59.6	62.35

Table 3. Parameters of rectangles R_1–R_3.

Rectangle	Parameters			
	Left	Right	Bottom	Top
R_1 in (X_9, X_8)	0.0020	0.1402	0.0734	0.1028
R_2 in (X_6, X_7)	0.7214	1.001	0.3484	1.001
R_3 in (X_6, X_7)	0.0081	0.6325	0.6014	1.001

addition do not go rectangles R_2 and R_3. (do not have node (x_6, x_7) in R_2 and R_3). Thus, Fig. 6 shows only graphs that satisfy Rule 1.

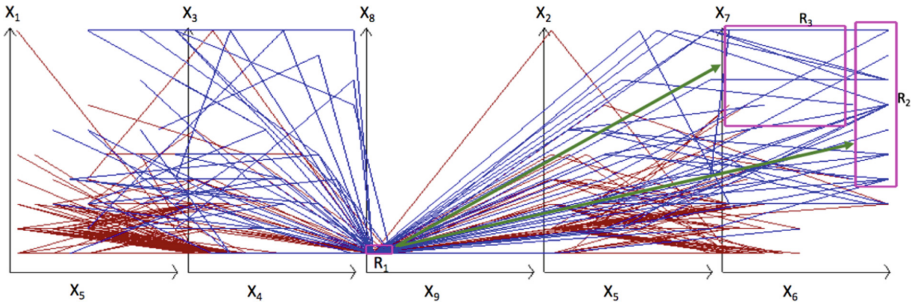


Fig. 5. WBC data in SPC as graphs representing 10-D points that go through R_1 without showing frequency of cases. Magenta boxes show rectangles R_1–R_3. (Color figure online)

In Fig. 5, the width of the lines are not adjusted their frequency, but in Fig. 6 the width of the lines is adjusted to their frequencies. The rule 1 covers 64.8% of all given 9-D points: 446 cases out of 688 cases (425 red cases and 21 blue cases) with the recall value 61.77% (425/688) as shown in Table 1. Among these 446 cases 303 Red and 17 Blue cases belong to training data and 122 Red and 4 blue cases belong to testing data.

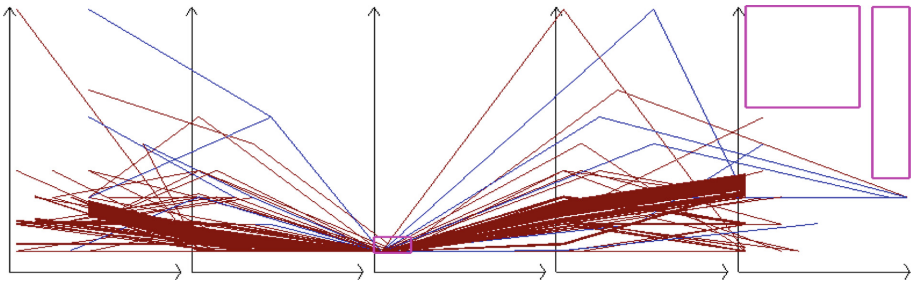


Fig. 6. WBC data in SPC as graphs representing 10-D points that satisfy Rule 1 (go through R_1 and not coming to R_2 and R_3 shown in magenta. Wide red lines show the frequency of red cases. (Color figure online)

The cases covered by rule 1 do not include only the 5.13% (23 cases) of the Red data, but include the 91.25% (219 cases) of the blue data. This rule covers only cases found in R_1 that do not come to R_2 or R_3. Rule 1 refused to predict other cases. Those other cases are dominantly blues (see Fig. 7) and must be covered by other rules. The WBC Rule 1 uses only 4 coordinates that form two pairs (X_9, X_8) and (X_6, X_7) therefore we can simplify the visualization of this rule by showing only them in SPC. It is done in Fig. 8 where each 4-D points is visualized losslessly as a single line. The advantage of this visualization is that it is easy to see and communicate to the medical

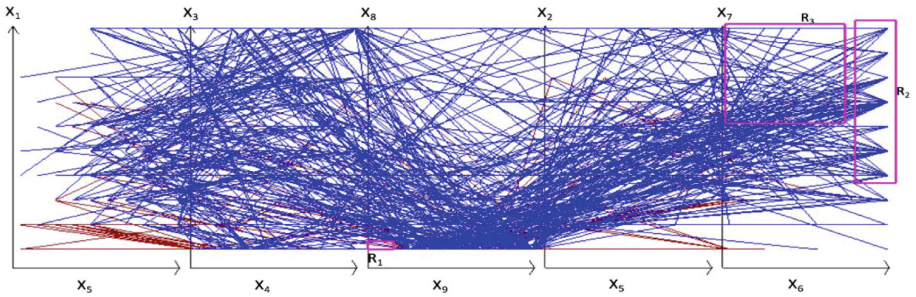


Fig. 7. Remaining WBC cases not covered by rule 1 (dominated by blue class). (Color figure online)

experts. The medical expert can easily understand this rule because it simply says that two attributes x_8 and x_9 must be in some limits identified in Table 3 and two other attributes x_6 and x_7 must not have values in some intervals that are shown visually in Fig. 8.

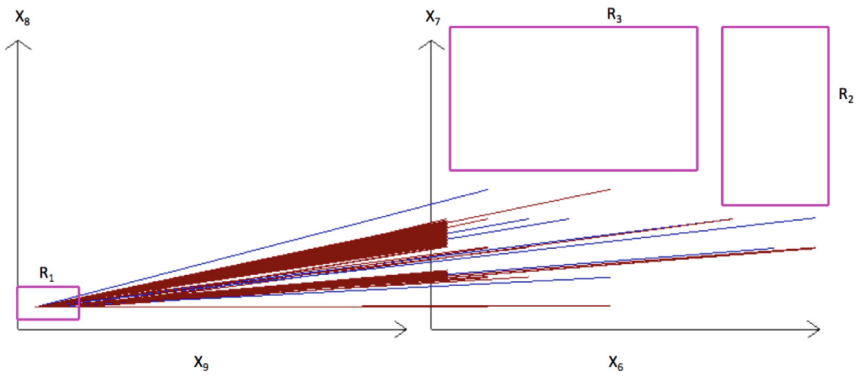


Fig. 8. WBC data in 4-D SPC as graphs in coordinates (X_9, X_8) and (X_6, X_7) that are used by Rule 1, i.e., WBC cases that go through R_1 and not go to R_2 and R_3 in these coordinates.

This allows the medical experts to analyze the consistency of this rule with the other medical domain knowledge, which is extremely difficult for the ML discrimination functions, which are "black" boxes or complex mathematical formulas.

The simple **WBC Rule 2** classifies all remaining cases (not covered by rule 1) to class 2 is

$$\begin{gathered}\text{If } (x_8, x_9) \in R_1 \,\&\, (x_6, x_7) \notin R_2 \,\&\, (x_6, x_7) \notin R_3 \\ \text{then } \mathbf{x} \in \text{class } 1(\text{Red, Benign}) \text{ else } \mathbf{x} \in \text{class } 2(\text{Blue, Malignant})\end{gathered} \tag{14}$$

This rule classifies all cases that are either in R_2 or in R_3 or not in R_1 as blue. This rule has accuracy 93.60% (425 + 219)/688) on *all 688 cases* as the confusion matrix

Table 4. Confusion matrixes of WBC rule 2.

Actual class	Confusion matrix on all 688 cases			Confusion matrix on 70% of training cases (482)			Confusion matrix on 30% on testing cases (206)		
	Predicted class			Predicted class			Predicted class		
	Red	Blue	Total	Red	Blue	Total	Red	Blue	Total
Red	425	23	448	303	19	322	122	4	126
Blue	21	219	240	17	143	160	4	76	80
Total	446	242	688	320	162	482	126	89	206

shows in Table 4. Its accuracy on *training data* is 92.53% and on *validation data* it is 96.11% computed from respective confusion matrixes shown in Table 4.

We found a better **WBC Rule 3** by searching the rectangles with highest density of blue cases:

$$\text{If } ((x_1, x_6) \in R_4 \vee (x_3, x_5) \in R_5) \,\&\, (x_2, x_5) \notin R_6 \text{ then } \mathbf{x} \in \text{class 2(Blue)} \quad (15)$$

This rule is of type of rules (4), but with the conclusion that **x** belongs to class 2, not class 1. Table 5 identifies the rectangles R_4–R_6 involved in this rule. WBC Rule 3 covers the 226 cases from class 2 (Blue), and the 12 cases from class 1 (Red) with the total precision of 94.95%. The **combined rule1 and rule 3** is as follows:

$$\text{If Rule1}(\mathbf{x}) \text{ then } \mathbf{x} \in \text{Red, else if Rule3}(\mathbf{x}) \text{ then x} \in \text{Blue, else refuse to classify } \mathbf{x} \quad (16)$$

The precision, recall and coverage of this rule relative to all cases are 95.18%, 94.62%, and 99.42% (Fig. 9). It is the performance details are in the confusion matrix in Table 6.

Table 5. Parameters of rectangles R_4–R_6 in normalized coordinates.

Rectangle	Parameters			
	Left	Right	Bottom	Top
R_4 in (X_1, X_6)	0.0010	0.9712	0.4180	1.001
R_5 in (X_3, X_5)	0.0000	0.9881	0.3154	0.7261
R_6 in (X_2, X_5)	0.0000	0.1003	0.143	0.3153

3.2 Experimental Case Study 2

The computational experiments with the 33-D Ionosphere data from the UCI Machine Learning repository [3] also show the efficiency of the FSP algorithm. To get the even number of coordinates and 17 pairs of coordinates, the algorithm in each epoch will select randomly the coordinate that will serve as coordinate X_{34}. In the following results X_{34} is a copy of X_{10}. The discovered patterns also were found by the search in the set of "rectangular" rules (1)–(8). In particular, on Ionosphere data, the FSP

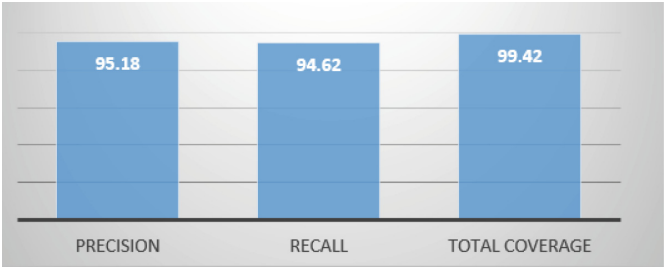


Fig. 9. The precision, recall and total coverage of combined rule 1 and rule 3

Table 6. Confusion matrix of combined rules 1 and 3 on all 688 cases.

	Predicted red class	Predicted blue class	Refusal	Total
Actual red class	425	12	2	437
Actual blue class	21	226	2	247
Total	446	238	4	688

algorithm found an efficient sequence of pairs of coordinates: (X_5, X_{26}), (X_{27}, X_{16}), (X_4, X_{11}), (X_{18}, X_{24}), (X_{28}, X_{31}), (X_{10}, X_3), (X_{23}, X_8), (X_{22}, X_{30}), (X_{21}, X_{10}), (X_{17}, X_2), (X_{15}, X_{33}), (X_{29}, X_{20}), (X_9, X_6), (X_{32}, X_{16}), (X_1, X_{25}), (X_{12}, X_{14}), (X_{19}, X_7). Here X_{10} is repeated in pairs (X_{10}, X_3) and (X_{21}, X_{10}). Similarly to the Case Study 1, the SPC visualization, with this sequence, reveals the visual classification pattern of *precision of over 90%* in all the 11 random 70%:30% splits of data into the training and the validation data (see Tables 7 and 8). The best precision on the training data is 98.36% with 100% precision on the validation data in one of these 70%:30% splits of the given data. The discovered rules in Tables 7 and 8 belong to the set of rules (4). The first rule in Table 7 that we denote as **Ionosphere Rule 1** is:

$$\begin{gathered}\text{If } [(x_4, x_{11}) \in R_1 \vee (x_5, x_{26}) \in R_8 \vee (x_{21}, x_{10}) \in R_9 \vee (x_{19}, x_7) \in R_{10}] \,\&\\ [(x_{27}, x_{16}) \notin R_3 \,\&\, (x_{28}, x_{31}) \notin R_6 \,\&\, (x_{23}, x_8) \notin R_4 \,\&\, (x_{17}, x_2) \in R_5 \,\&\, (x_{15}, x_{33}) \notin R_7 \,\&\\ (x_9, x_6) \notin R_2]\\ \text{then x} \in \text{class 1(Red, Good)}\end{gathered} \tag{17}$$

where R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, and R_{10} are specific rectangles, in respective pairs of Cartesian coordinates (X_4, X_{11}), (X_9, X_6), (X_{27}, X_{16}), (X_{23}, X_8), (X_{17}, X_2), (X_{28}, X_{31}), (X_{15}, X_{33}), (X_5, X_{26}), (X_{21}, X_{10}), and (X_{19}, X_7). Table 9 shows the parameters of R_1–R_{10} in the normalized coordinates. This rule, with a random 70%:30% data split into the training and validation data, has the precision 98.36% on the training data and 100% on validation data. Figures 10 and 11 show its rectangles R_1–R_{10} in the SPCs as magenta boxes and cases that satisfy rule 1 in Fig. 11 and all cases in Fig. 10. Figure 12 shows the remaining cases.

Table 7. Number of cases that satisfy the Ionosphere Rule 1 in 11 random 70%:30% splits of data.

70%:30% random data splits	Number of cases that satisfy if part of the Rule 1					
	Red class			Blue class		
	Training	Testing	Total	Training	Testing	Total
1	180	45	225	3	0	3
2	172	52	224	8	2	10
3	133	92	225	7	3	10
4	129	95	224	6	2	8
5	200	24	224	9	0	3
6	160	65	225	10	2	12
7	183	42	225	11	1	13
8	158	67	225	11	2	13
9	191	34	225	13	1	14
10	184	37	221	7	0	7
11	157	66	223	12	2	14
Average	170	54	224	9	1	9

Table 8. Precision, recall and coverage of Ionosphere Rule 1 in 11 random 70%:30% splits of data.

70%:30% random data splits	Rule precision		Rule recall (correct) coverage			Rule total coverage of cases, %
	Training, %	Testing, %	Training, %	Testing, %	Total, %	
1	98.36	100	51.28	12.82	64.1	64.95
2	95.55	96.29	49	14.81	63.81	66.6
3	95	96.84	37.89	26.21	64	66.95
4	95.5	97.93	36.75	27.06	63.81	66.09
5	95.69	100	56.98	6.83	63.81	64.67
6	94.1	97.01	45.58	18.51	64.1	67.52
7	94.32	97.67	52.13	11.96	64.1	67.80
8	93.491	97.1	45.01	19.08	64.09	67.80
9	93.62	97.14	54.41	9.68	64.1	68.09
10	96.33	100	52.421	10.54	62.96	64.95
11	92.89	97.05	44.72	18.8	63.53	67.52
Average	94.99	97.93	48.43	15.38	63.81	66.38

The Rule 1 uses only 20 coordinates that form 10 pairs (X_5, X_{26}), (X_{27}, X_{16}), (X_4, X_{11}), (X_{28}, X_{31}), (X_{23}, X_8), (X_{21}, X_{10}), (X_{17}, X_2), (X_{15}, X_{33}), (X_9, X_6) and (X_{19}, X_7). Therefore, we simplify its visualization by showing only them in the SPCs (see Figs. 13, 14 and 15) with the lossless visualization of each of the 20-D points, as a single polyline. The simple **Ionosphere Rule 2** classifies all remaining cases (not covered by rule 1) to class 2 is

Table 9. Parameters of rectangles R_4–R_6 in normalized coordinates.

Rectangle	Parameters			
	Left	Right	Bottom	Top
R_1 in (X_4, X_{11})	0.23912	1.001	0.663667	0.999667
R_2 in (X_9, X_6)	0.1979	0.99869	0	0.0403333
R_3 in (X_{27}, X_{16})	0.130225	1.001	0	0.0256667
R_4 in (X_{23}, X_8)	0.29087	0.58943	1.001	0.946
R_5 in (X_{17}, X_2)	0.40433	0.9998	0	0.022
R_6 in (X_{28}, X_{31})	0.30765	1.001	0	0.0366667
R_7 in (X_{15}, X_{33})	0.25733	1.001	0.960667	0.998667
R_8 in (X_5, X_{26})	0.125675	0.575667	0.575667	0.645333
R_9 in (X_{21}, X_{10})	0.25278	0.50868	0.0476667	0.0843333
R_{10} in (X_{19}, X_7)	0.4203	0.8254	0.194333	0.432667

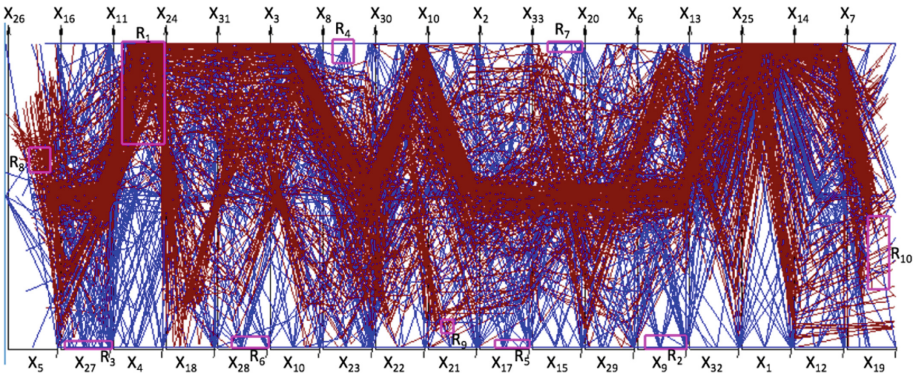


Fig. 10. 351 Ionosphere cases in the SPCs as graphs of 34-D points (good cases in red and bad cases in blue). Rectangles that are used in Ionosphere Rule 1 are in magenta. (Color figure online)

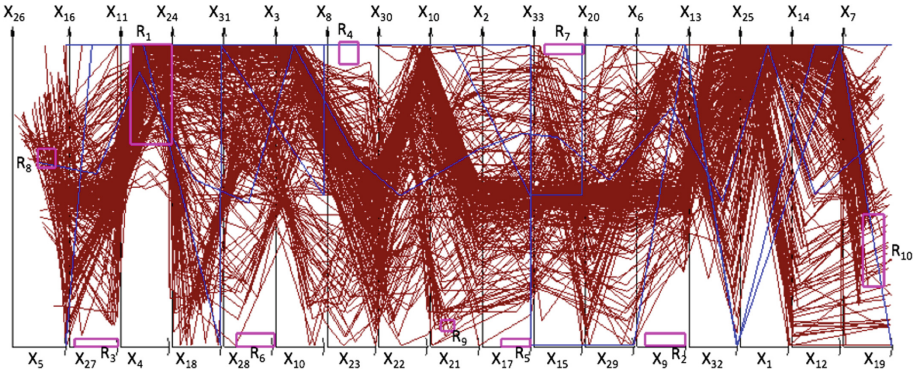


Fig. 11. 34-D Ionosphere cases covered by Ionosphere rule 1. Rectangles from rule 1 are in magenta. (Color figure online)

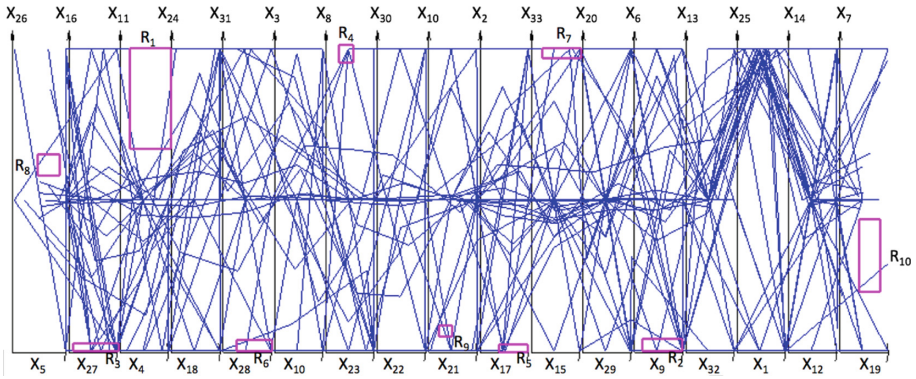


Fig. 12. Remaining Ionosphere cases (cases not covered by Ionosphere rule 1).

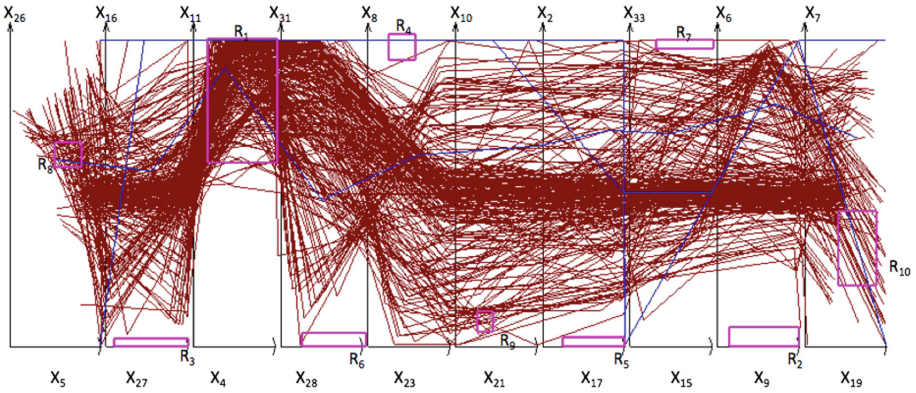


Fig. 13. Ionosphere cases in 20-D points covered by the Ionosphere rule 1.

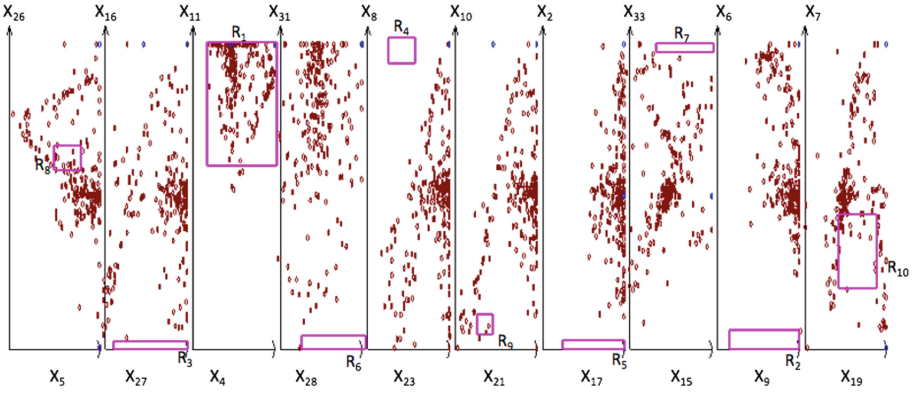


Fig. 14. Ionosphere cases in 20-D points covered by rule 1.

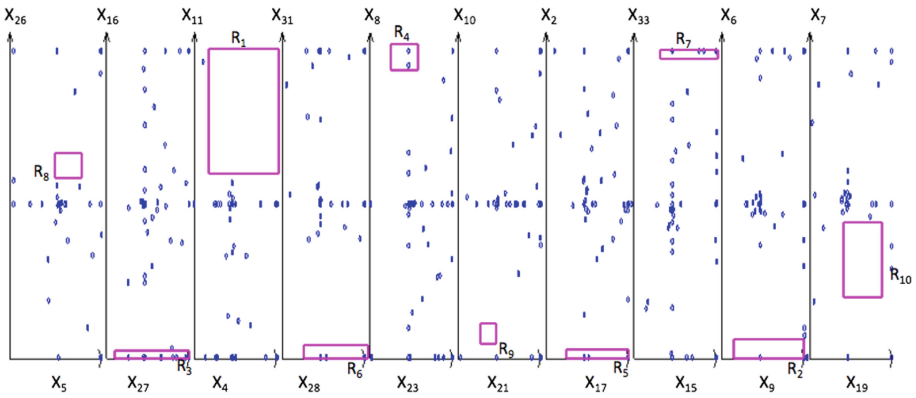


Fig. 15. Ionosphere cases in 20-D points not covered by rule 1.

$$\begin{gathered}\text{If } [(x_4, x_{11}) \in R_1 \vee (x_5, x_{26}) \in R_8 \vee (x_{21}, x_{10}) \in R_9 \vee (x_{19}, x_7) \in R_{10}] \,\&\\ [(x_{27}, x_{16}) \notin R_3 \,\&\, (x_{28}, x_{31}) \notin R_6 \,\&\, (x_{23}, x_8) \notin R_4 \,\&\, (x_{17}, x_2) \notin R_5 \,\&\, (x_{15}, x_{33}) \notin R_7 \,\&\\ (x_9, x_6) \notin R_2]\\ \text{then x} \in \text{class 1(Red, Good) else x} \in \text{class 2(Blue, Bad)}\end{gathered} \tag{18}$$

This rule classifies all the cases rejected by rule 1 as blue with 99.14% accuracy on *all the 351 cases* (225 + 123)/351), see the confusion matrix in Table 10. Its accuracy on *training data* is 98.78% and 100% on the *validation data* based on the confusion matrixes in Table 10.

Table 10. Confusion matrixes of the Ionosphere rule 2.

Actual class	Confusion matrix on all 351 cases			Confusion matrix on training data (246, 70%)			Confusion matrix on validation data (206, 30%)		
	Predicted class			Predicted class			Predicted class		
	Red	Blue	Total	Red	Blue	Total	Red	Blue	Total
Red	225	3	228	180	3	183	45	0	45
Blue	0	123	123	0	63	63	0	60	60
Total	225	126	351	180	66	246	45	60	105

3.3 Experimental Case Study 3

The computational experiments with the 8-D Abalone data, male and infant cases, from the UCI Machine Learning repository [3] also show the efficiency of the FSP algorithm. The discovered patterns were found by the search in the set of "rectangular" rules (1)–(8). In particular, the FSP algorithm found an efficient sequence of coordinate pairs: (X_5, X_6), (X_1, X_2), (X_3, X_8), (X_4, X_7). Similarly, to Case Studies 1 and 2, the

SPC visualization, with this sequence, reveals the visual classification pattern of with the *precision of over 90%* (see Tables 11 and 12) in the 11 random 70%:30% splits of data into the training and the validation cases. The best precision is 92.06% on the training data and 96.17% on the validation data. Figure 16 shows 2870 Abalone data in the SPCs, as graphs of 8-D points, with the male cases in red, and the infant cases in blue. Rectangles used in Abalone Rule 1 defined below are in magenta. Figures 17 and 18 show cases covered this rule and Fig. 19 shows the remaining cases.

Table 11. Number of cases satisfying the if part of Abalone Rule 1 in 11 random 70%:30% splits.

70%:30% random data splits	Number of cases that satisfy if part of the Rule 1					
	Red class			Blue class		
	Training	Testing	Total	Training	Testing	Total
1	1193	304	1497	103	12	115
2	1034	396	1430	103	37	140
3	972	463	1435	86	46	132
4	1015	484	1499	105	40	145
5	1224	272	1496	126	10	136
6	1003	371	1374	98	28	126
7	989	402	1391	92	37	129
8	1077	389	1466	108	22	130
9	1146	328	1474	117	31	148
10	1201	276	1477	121	15	136
11	996	357	1353	103	35	138
Average	1078	367	1445	106	29	134

The discovered rules shown in the Tables 11 and 12 belong to the set of rules (4). The first rule in the Table 11, which we denote as the **Abalone Rule 1** is:

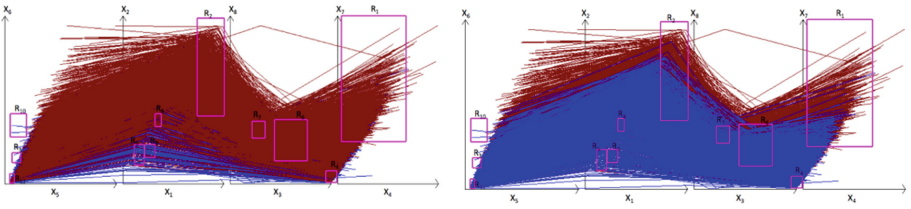
$$\text{If } [(x_4,x_7) \in R_1 \vee (x_1,x_2) \in (R_2 \vee R_8) \vee (x_3,x_8) \in (R_4 \vee R_6)] \,\&\, [(x_5,x_6) \notin (R_3 \,\&\, R_{10} \,\&\, R_{11}) \,\&\, (x_1,x_2) \notin (R_5 \,\&\, R_9) \,\&\, (x_3,x_8) \in (R_4 \,\&\, R_7)], \text{ then } \mathbf{x} \in \text{class 1}(\text{Red}, \text{Male}), \tag{19}$$

where R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, $R_{10,}$ and R_{11} are specific rectangles (see Table 13), in respective pairs of the Cartesian coordinates (X_4, X_7), (X_1, X_2), (X_5, X_6), (X_3, X_8), (X_1, X_2), (X_3, X_8), (X_3, X_8), (X_1, X_2), (X_1, X_2), (X_5, X_6), and (X_5, X_6). The simple **Abalone Rule 2,** which classifies all the remaining cases (not covered by rule 1) into class 2 is:

$$\begin{gathered}\text{If } [(x_4,x_7) \in R_1 \vee (x_1,x_2) \in (R_2 \vee R_8) \vee (x_3,x_8) \in (R_4 \vee R_6)] \,\&\, [(x_5,x_6) \notin (R_3 \,\&\, R_{10} \& R_{11}) \\ \& (x_1,x_2) \notin (R_5 \,\&\, R_9) \,\&\, (x_3,x_8) \notin (R_4 \,\&\, R_7)] \\ \text{then } \mathbf{x} \in \text{class 1}(\text{Red}, \text{Male}), \text{ else } \mathbf{x} \in \text{class 2}(\text{Blue}, \text{Infant})\end{gathered} \tag{20}$$

Table 12. Precision, recall and coverage of Rule 1 in 11 random 70%:30% splits of data.

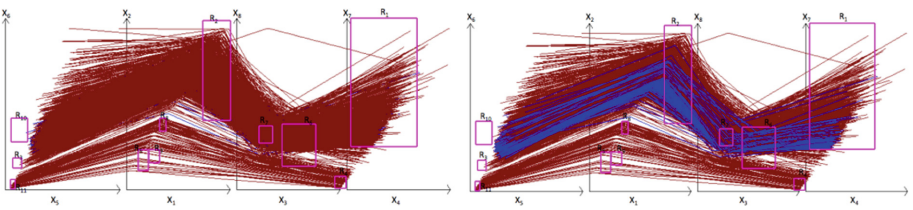
70%:30% random data splits	Rule precision		Rule recall (correct) coverage			Rule total coverage of cases, %
	Training, %	Testing, %	Training, %	Testing, %	Total, %	
1	92.05	96.20	41.56	10.59	52.16	56.16
2	90.94	91.45	36.02	13.79	49.82	54.70
3	91.87	90.96	33.86	16.13	50	54.59
4	90.62	92.36	35.36	16.86	52.22	57.28
5	90.66	96.45	42.64	9.47	52.12	56.86
6	91.09	92.98	34.94	12.92	47.87	52.26
7	91.48	91.57	34.45	14.00	48.46	52.96
8	90.88	94.64	37.52	13.55	51.08	55.6
9	90.73	91.36	39.93	11.42	51.35	56.51
10	90.84	94.84	41.84	9.616	51.46	56.2
11	90.62	91.07	34.70	12.43	47.14	51.95
Average	91.07	93.06	37.53	12.80	50.33	55.01



(a) Red cases visualized on the top blue cases.

(b) Blue cases visualized on the top the red cases.

Fig. 16. A set of 2870 abalone data visualized in the SPCs, as graphs of 8-D points, with the male cases in red, and the infant cases in blue. Rectangles used in Abalone Rule 1 are in magenta. (Color figure online)



(a) Red cases on the top of the blue cases.

(b) blue cases on the top of the red cases.

Fig. 17. 8-D abalone cases covered by abalone Rule 1. (Color figure online)

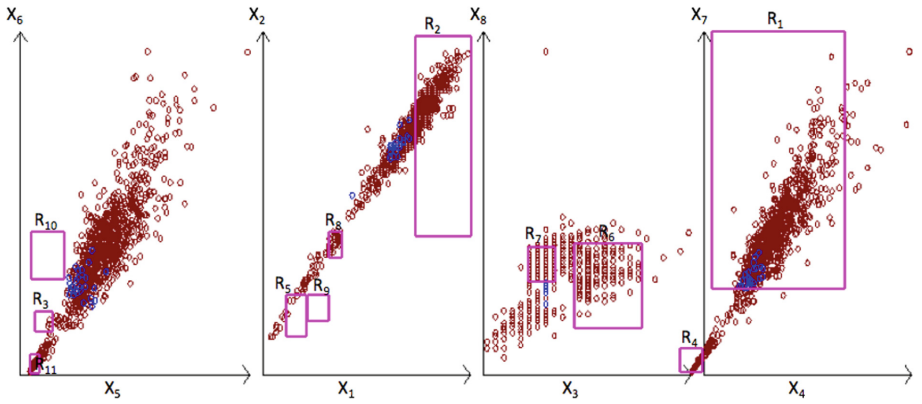
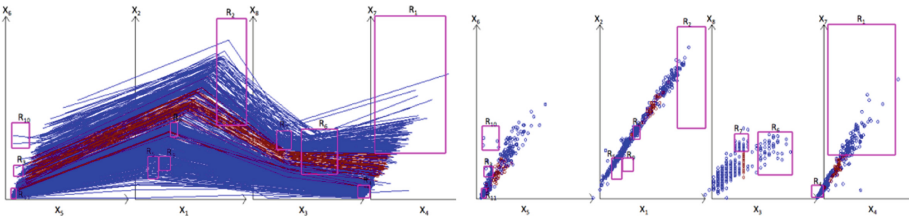


Fig. 18. 8-D Abalone cases covered by rule 1 (only nodes of graphs are shown to decrease occlusion)



(a) Full graphs. (b) Only nodes of graphs.

Fig. 19. Remaining Abalone cases (cases not covered by Abalone rule 1).

Table 13. Parameters of rectangles R_1–R_{11}.

Rectangle	Parameters			
	Left	Right	Bottom	Top
R_1 in (X_4, X_7)	0.09068	0.6946	0.260333	1.00
R_2 in (X_1, X_2)	0.7458	0.99955	0.421667	0.99996
R_3 in (X_5, X_6)	0.0209	0.1015	0.128333	0.190667
R_4 in (X_3, X_8)	0.94857	1.001	0	0.077
R_5 in (X_1, X_2)	0.162175	0.253525	0.242	0.113667
R_6 in (X_3, X_8)	0.46645	0.77602	0.139333	0.399667
R_7 in (X_3, X_8)	0.25837	0.38017	0.282333	0.388667
R_8 in (X_1, X_2)	0.355025	0.415925	0.355667	0.436333
R_9 in (X_1, X_2)	0.26	0.355025	0.161333	0.242
R_{10} in (X_5, X_6)	0.015075	0.167325	0.289667	0.436333
R_{11} in (X_5, X_6)	0.012	0.035675	0.0	0.0586667

with accuracy 94.91% on *all cases*, 93.33% and 98.60% on *training* and *validation data* based on the confusion matrixes Table 14.

Table 14. Confusion matrixes of Abalone Rule 2.

Actual class	Confusion matrix on all 2870 cases			Confusion matrix on training cases (2009, 70%)			Confusion matrix on validation cases (861, 30%)		
	Predicted class			Predicted class			Predicted class		
	Red	Blue	Total	Red	Blue	Total	Red	Blue	Total
Red	1497	115	1612	1193	103	1296	304	12	316
Blue	31	1227	1258	31	682	713	0	545	545
Total	1528	1342	2870	1224	785	2009	304	557	861

4 Comparison with Published Results and Conclusion

Case Study 1: the best accuracy reported for the Wisconsin Breast Cancer (WBC) dataset for the SVM in [5] is 96.995% with the 10-fold cross-validation tests. Other results are 96.84% [6] and 96.99% [7] for the SVM, and 97.28% [5] by combing SVM, C4.5 decision tree, naïve Bayesian classifier, and the k-Nearest Neighbors algorithms.

These models classify all the cases, while many of our rules refuse to classify some of the cases. Our WBC rule 2 classify all cases, but with lower accuracy 93.60%. Our better rule that combines WBC rules 1 and 3 has precision 95.18%. While in general, accuracy and precision are different, here the combination of rules 1 and 3 cover almost all cases (99.42%, only 4 cases are refused). The precision for such high coverage is almost identical to accuracy. Thus, it is quite close to the published results, but slightly lower. However, in contrast with SVM, it is *visual*, has clear *interpretation* and *explainable to a domain expert* which is very important in domains with high cost of errors where the explanation of the model is mandatory.

Case Study 2: for Ionosphere dataset, the highest accuracy reported by [8] is 98% on training and 93% on validation data using the multilayer perceptron. Other results are 94.87% by using C4.5 algorithm and 94.59% using Rule Induction RIAC algorithm [9], 97.33% by SVM with Particle Swarm Optimization and 10-fold cross-validation [10].

These models classify all the objects, while many of our rules refuse to classify some of the cases. In contrast, our Ionosphere rule 2 classify all cases. For this rule, precision is identical to accuracy that is 98.78% on training data and 100% on validation data. Thus, our results are slightly higher, than those for the published classification models.

Case Study 3: The highest accuracy reported in [11] for Abalone dataset using SVM is 99.26% with 5-fold Cross-validation for all three classes. Another result is 97.80% accuracy using a case base reasoning method [12]. Our result, that are within [93.33,

98.60]% interval, are quite close to these published results. Unlike [11], we use a more challenging for classification 70:30 split, than the 5-fold that is 80:20 split.

Our Ionosphere and Abalone rules 2 are also *visual*, *interpretable,* and *explainable* to a domain expert, which is critical, in many domains with mandatory model explanation. This comparison shows that the proposed FSP algorithm, with the SPC visualization, produced the results comparable with the other major machine learning algorithms, in the accuracy and precision. The FSP algorithm has the following significant advantages: it is (i) *visual* with *minimal occlusion*, (ii), *interactive*, (iii) *understandable* by the user, and (iv) *simpler* than many machine-learning algorithms. The future study is using more interpretable rules for discovery by the FSP algorithm along with the other General Line Coordinates, beyond SPC.

References

1. Kovalerchuk, B., Grishin, V.: Adjustable general line coordinates for visual knowledge discovery in n-D data. Inf. Vis. (2017). https://doi.org/10.1177/1473871617715860
2. Kovalerchuk, B.: Visual Knowledge Discovery and Machine Learning. Springer, Heidelberg (2018). https://doi.org/10.1007/978-3-319-73040-0
3. Lichman, M.: UCI Machine Learning Repository (2013). http://archive.ics.uci.edu/ml
4. Wilinski, A., Kovalerchuk, B.: Visual knowledge discovery and machine learning for investment strategy. Cogn. Syst. Res. **44**, 100–114 (2017)
5. Salama, G.I., Abdelhalim, M., Zeid, M.A.: Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) **32**, 2 (2012)
6. Aruna, S., Rajagopalan, D.S., Nandakishore, L.V.: Knowledge based analysis of various statistical tools in detecting breast cancer. Comput. Sci. Inf. Technol. **2**, 37–45 (2011)
7. Christobel, A., Sivaprakasam, Y.: An empirical comparison of data mining classification methods. Int. J. Comput. Inf. Syst. **3**, 24–28 (2011)
8. Duch, W., Kordos, M.: Multilayer perceptron with numerical gradient. In: ICANN 2003, pp. 106–109 (2003)
9. Hamilton H.J., Shan N., Cercone N., RIAC: A Rule Induction Algorithm Based on Approximate Classification. Computer Science Department, University of Regina (1996)
10. Tu, C.-J., Chuang, L.Y., Yang, C.H.: Feature selection Using PSO-SVM. IAENG Int. J. Comput. Sci. **33**(1), 111–116 (2007)
11. Sain, H., Purnami, S.W.: Combine sampling support vector machine for imbalanced data classification. Procedia Comput. Sci. **1**(72), 59–66 (2015)
12. Smiti, A., Elouedi, Z.: Maintaining case based reasoning systems based on soft competence model. In: Polycarpou, M., de Carvalho, A.C.P.L.F., Pan, J.S., Woźniak, M., Quintian, H., Corchado, E. (eds.) HAIS 2014. LNCS (LNAI), vol. 8480, pp. 666–677. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07617-1_58