

Feature Transformation Methods in Data Mining

Andrew Kusiak, *Member, IEEE*

Abstract—The quality of knowledge extracted from a data set can be enhanced by its transformation. Discretization and filling missing data are the most common forms of data transformation. A new transformation method named feature bundling is introduced. A feature bundle involves a set of features in its pure or transformed form. The computational results reported in this paper show that the classification accuracy of decision rules generated from data sets with feature bundles is enhanced. The proposed concept of feature bundling is applied to a data set from semiconductor industry.

Index Terms—Classification, data mining, decision making, feature bundling, feature transformation method, knowledge discovery, transformed data set.

I. INTRODUCTION

THE recent advances in data mining have produced algorithms for extracting knowledge contained in large data sets. This knowledge can be explicit, e.g., represented as decision rules, and utilized for decision making in areas where decision models do not exist. The machine learning algorithms construct associations among various parameters (called features in data mining and attributes in computer database literature) that are important in modeling processes.

Learning (classification) systems of interest to this research fall into eight general categories.

- A) Classical statistical methods (e.g., linear discriminant, quadratic discriminant, and logistic discriminant analyses) [1].
- B) Modern statistical techniques (e.g., projection pursuit classification, density estimation, k -nearest neighbor, casual networks, Bayes theorem [2]).
- C) Neural networks (e.g., backpropagation, Kohonen, linear vector quantifiers, and radial function networks) [1].
- D) Vector support machines [3].
- E) Decision tree methods (e.g., ID3 [4], CN2 [5], C4.5 [6], T2 [7], Lazy decision trees [8], OODG [9], OC1 [10], AC, BayTree, CAL5, CART, ID5R, IDL, TDIDT, and PROSM; see [1] for the description of these algorithms).
- F) Decision rule algorithms (e.g., AQ15 [11], [12], LERS [13] and numerous other algorithms based on the rough set theory [14] and [15]).
- G) Learning classifier systems (e.g., GOFFER-1 [16], MonaLysa [17], and XCS [18] and [19]).
- H) Association rule algorithms (e.g., DB2IntelligentMiner) [19].

TABLE I
DATA SET

No.	F1	F2	F3	F4	D
1	Red	1.2	2	0	Low
2	Blue	6.1	2	2	High
3	Red	4.2	1	0	Low
4	Yellow	1.2	0	1	Medium
5	Green	3.8	3	1	Low
6	Yellow	8.8	1	0	Medium
7	Red	5.4	0	2	High

Lim [20] presented a comprehensive comparative study of over thirty learning algorithms of categories A, B, C, E, and F. The background of the category D learning algorithms is provided in [21]. The algorithms of class G are discussed in [22]. The roots of the class G algorithms are in evolutionary computation initiated by Holland [23] and Goldberg [24]. Class H and many other algorithms are discussed in [25].

In this paper, the decision rule algorithms (category F) will be explored for two major reasons.

- 1) Generation of explicit knowledge in the form acceptable by a user: User is able to understand the extracted knowledge, assess its usefulness, and learn new and interesting concepts.
- 2) Controllable classification accuracy: This characteristic is due to the nature of the algorithms themselves as well as the feature transformation concept discussed in this paper.

The concepts presented in the paper are general and applicable to data sets mined by many other algorithms.

Consider the data set in Table I with seven objects, four features, and decision D .

The rules extracted by a decision rule algorithm are shown in Fig. 1. The numbers behind each rule correspond to the row numbers in Table I (for more details see Section III).

The decision rules of Fig. 1 are represented as the graph of Fig. 2.

The pattern corresponding to this rule tree is shown in Table II.

The pattern in Table II and the corresponding tree in Fig. 2 use only three features F1, F2, and F4 to represent all objects (rows) of Table I.

II. FEATURE TRANSFORMATION METHODS

Data sets can be mined in their raw collected form, or they can be transformed. The following transformation methods can be applied to data sets:

- a) filling in missing values;
- b) discretization;

Manuscript received February 14, 2001; revised July 25, 2001.

The author is with the Intelligent Systems Laboratory, Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA 52242-1527 USA (e-mail: andrew-kusiak@uiowa.edu).

Publisher Item Identifier S 1521-334X(01)08957-1.

```

Rule1. IF (F1 = {Green OR Red}) AND (F2 < 4.8) THEN (D = Low); [1, 3, 5]
Rule2. IF (F1 = Yellow) THEN (D = Medium); [4, 6]
Rule3. IF (F4 = 2) THEN (D = High); [2, 7]
    
```

Fig. 1. Decision rules derived by a rough set algorithm.

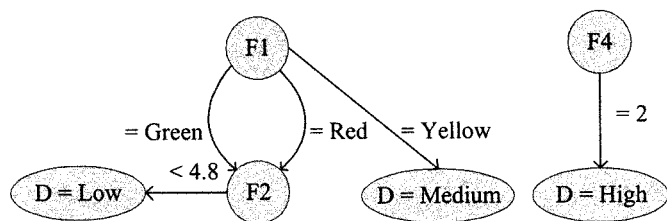


Fig. 2. Rule tree derived from the data set in Table I.

TABLE II
PATTERN CORRESPONDING TO THE RULE TREE IN FIG. 2

No	F1	F2	F3	F4	D
1	Red	1.2	2	0	Low
2	Blue	6.1	2	2	High
3	Red	4.2	1	0	Low
4	Yellow	1.2	0	1	Medium
5	Green	3.8	3	1	Low
6	Yellow	8.8	1	0	Medium
7	Red	5.4	0	2	High

TABLE III
TRAINING DATA SET

No.	F1	F2	F3	F4	D
1	0	1	0	2	Zero
2	1	1	0	2	Two
3	0	0	0	1	Zero
4	0	1	1	0	One
5	0	0	1	3	Zero

TABLE IV
CLASSIFICATION QUALITY OF THE FEATURES IN TABLE III

Classification quality	F1	F2	F3	F4
	.2	.4	0.	.6

Feature bundling is of particular interest in temporal data mining as relationships are formed among features rather than their values. Such relationships tend to be more stable in time comparing to the relationships among feature values. Although feature bundling is primarily intended for integer, normative and categorical features, it can be extended to features with continuous values, for example, by using regression functions.

- c) feature content modification (generalization, specialization);
- d) feature bundling.

In this paper, the fourth method, the feature bundling method, is introduced. Rather than concentrating on individual features, the approach presented in this paper advocates capturing relationships among feature bundles. A feature bundle is a collection of features in its pure or transformed form.

The following feature bundles are proposed.

- 1) $F5 \langle \text{logical operator} \rangle F2 \langle \text{logical operator} \rangle F9$, where the $\langle \text{logical operator} \rangle = \{\text{AND, OR, NOT, EXOR}\}$. Note that an ordered set of features linked by the AND operator becomes a sequence, e.g., $F2 \text{ AND } F9 \text{ AND } F4$ is denoted as $F2.F9.F4$.
- 2) $F3 \langle \text{arithmetic operator} \rangle F8$.
- 3) $aF2 \langle \text{arithmetic operator} \rangle bF4 / \langle \text{arithmetic operator} \rangle cF5$, where a , b , and c are constants and the $\langle \text{arithmetic operator} \rangle = \{+, -, /, \times\}$.
- 4) Regression function defined on a subset of features.

Of interest to data mining are bundles that form time-invariant relationships. These relationships are expressed by comparators such as \geq , \leq , $>$, $<$, $=$, \neq , \approx , ∞ . For example, the two features $F3$ and $F7$ can be described as exponential functions, however, their relationship is $F3 < F7$ for $F9 \in [5, 16]$ and $F3 \geq F7$ for $F9 \notin [5, 16]$ as stated by the following two rules:

```

IF  $F9 \in [5, 16]$  AND  $F3 < F7$  THEN  $D = \text{Hot}$ 
IF  $F9 \notin [5, 16]$  AND  $F3 \geq F7$  THEN  $D = \text{Cold}$ 
    
```

III. TRANSFORMED DATA SETS

In this section the feature bundling concept is illustrated with mining a data set involving feature sequences. Once the feature sequences are defined, the original data set is transformed according to the newly created features and it is used for mining. The reason for forming feature sequences is threefold:

- 1) improved classification accuracy (defined in the Appendix) of decision rules with feature sequences; The classification accuracy gain will be demonstrated later in this section;
- 2) collective judgement based on a selected set of features; e.g., a gene sequence;
- 3) some features are generated in clusters that may naturally form bundles, e.g., process control parameters.

The analogy of a feature sequence to a gene sequence appears to have merits. The increased classification accuracy of rules with feature sequences may be traced to analogy with functionality and expression of gene sequences.

The impact of feature sequences on classification accuracy is discussed next.

Consider the data set in Table III consisting of five objects (rows) each with four features $F1-F4$ and decision D .

The term classification quality (defined in the Appendix) is used to analyze the properties of features of a data set.

The values of classification quality (CQ) of the features in Table III are shown in Table IV.

```

Rule 1. IF (F2 = 0) THEN (D = Zero); [2, 66.67%, 100.00%] [3, 5]
Rule 2. IF (F1 = 0) AND (F3 = 0) THEN (D = Zero); [2, 66.67%, 100.00%] [1, 3]
Rule 3. IF (F4 = 0) THEN (D = One); [1, 100.00%, 100.00%] [4]
Rule 4. IF (F1 = 1) THEN (D = Two); [1, 100.00%, 100.00%] [2]

```

Fig. 3. Rules derived from the data set in Table III.

TABLE V
ABSOLUTE CLASSIFICATION ACCURACY FOR THE DATA SET IN TABLE III

D	Zero	One	Two	None
Zero	1	1	0	1
One	1	0	0	0
Two	1	0	0	0

TABLE VI
CLASSIFICATION ACCURACY FOR THE DATA SET IN TABLE III

	Correct	Incorrect	None
Zero	33.33%	33.33%	33.33%
One	0.00%	100.00%	0.00%
Two	0.00%	100.00%	0.00%
Av	20.00%	60.00%	20.00%

The illustrative rule set derived from the data set in Table III is shown in Fig. 3.

These decision rules in Fig. 3 are presented in the following format: IF (Condition) THEN (Outcome); [Rule support, Relative rule strength, Discrimination level] [Objects represented by the rule] (For the definitions of these terms see the Appendix).

For example, Rule 2

IF (F1 = 0) AND (F3 = 0) THEN (D = Zero);
[2, 66.67%, 100.00%][1, 3] reads

IF (The value of feature F1 equals 0) AND (The value of F3 equals 0) THEN (The decision D is Zero); [This rule represents 2 objects; In this case these two objects make up 66.67% of all objects in the training data set with the decision $D = \text{Zero}$; The two objects exactly match the conditions and decision of this rule] [The objects represented by this rule are 1 and 3].

The quality of predictions with the rules extracted from a data set is usually evaluated by a cross-validation scheme [26]. Due to the small size of the data set in Table III, the 1-out-of- n ($n = 5$) cross-validation scheme was used. An object was removed from the training data set, one at a time, and the rules were extracted from the set of $n - 1 = 4$ objects. They were in turn used to predict decision D of the previously removed object. This process was repeated $n = 5$ times. The testing has produced the absolute classification results in Tables V and VI.

The diagonal numbers in the matrix of Table V (called also a confusion matrix) represent the number of outcomes D that have been correctly classified. In this case 1 of the 3 decisions $D = \text{Zero}$ have been correctly predicted, 0 decisions $D = \text{One}$ and 0 decisions of $D = \text{Two}$ have been correctly classified. The numbers off the diagonal indicate the incorrectly predicted decisions, e.g., for decision $D = \text{Zero}$ (row Zero in Table V) one decision $D = \text{One}$ was generated and one object could not be classified, i.e., it was assigned to the "None" category.

TABLE VII
DATA SET WITH FEATURE SEQUENCE F2_F3

No.	F1	F2	F3	F4	D
1	0	1	0	2	0
2	1	1	0	2	2
3	0	0	0	1	0
4	0	1	1	0	1
5	0	0	1	3	0

TABLE VIII
CLASSIFICATION QUALITY OF THE FEATURES IN TABLE VII

Classification quality	F1	F2	F3	F4
	.2	.6	.6	

Table VI reports the percentage of objects that have been classified correctly, incorrectly, or fall into the "None" category for the three decision values $D = \text{Zero}$, One , or Two .

The last row in Table VI includes the average classification accuracy for the rules extracted from $n - 1 = 4$ objects for $n = 5$.

Rather than directly extracting rules from a data set, in this paper data sets with feature sequences are considered. The original data set is transformed into a data set where the feature values corresponding to the feature sequences are merged. The transformed data set is used for rule extraction. The data set with the feature sequence F2_F3 is shown in Table VII.

The classification quality of the features in Table VII is provided in Table VIII.

The classification quality of the feature sequence F2_F3 is considerably higher than the classification quality of the component features F2 and F3 reported in Table IV.

Fig. 4 presents the rules extracted from the data set in Table VII.

The 1-out-of- n ($n = 5$) cross validation results for the rules in Fig. 4 are reported in Tables IX and X. It should be noted that the decision $D = \text{Zero}$ for each of the three objects has been correctly predicted (100% classification accuracy in the class $D = \text{Zero}$), as opposed to the classification accuracy of 33.33% reported in Table VII or the correctly predicted decision for one object in Table V.

The classification quality of the individual features F1-F4 and the feature sequences F1_F3, F1_F4, F2_F3, F2_F4, and F3_F4 is shown in Table XI. The value of the classification quality (CQ) varies from 0 for feature F3 and 1 for the feature sequence F1_F4.

Table XII summarizes the 1-out-of- n (for $n = 5$) validation results for nine data sets, the data set in Table III and eight data sets with one or more feature sequences. Note that in the fea-

```

Rule 5. IF (F2_3 = 0_0) THEN (D = Zero); [1, 33.33%, 100.00%] [3]
Rule 6. IF (F2_3 = 0_1) THEN (D = Zero); [1, 33.33%, 100.00%] [5]
Rule 7. IF (F1 = 0) AND (F4 = 2) THEN (D = Zero); [1, 33.33%, 100.00%] [[1]
Rule 8. IF (F2_3 = 1_1) THEN (D = One); [1, 100.00%, 100.00%] [4]
Rule 9. IF (F1 = 1) THEN (D = Two); [1, 100.00%, 100.00%] [2]
    
```

Fig. 4. Rules derived from the data set in Table VII.

TABLE IX
ABSOLUTE CLASSIFICATION ACCURACY FOR THE DATA SET IN TABLE VII

D	Zero	One	Two	None
Zero	3	0	0	0
One	1	0	0	0
Two	1	0	0	0

TABLE X
CLASSIFICATION ACCURACY FOR THE DATA SET IN TABLE VII

	Correct	Incorrect	None
0	100.00%	0.00%	0.00%
1	0.00%	100.00%	0.00%
2	0.00%	100.00%	0.00%
Av	60.00%	40.00%	0.00%

TABLE XI
SUMMARY OF FEATURE CLASSIFICATION QUALITY

Feature set	F1	F2	F3	F4	F1_F2	F1_F3	F1_F4	F2_F3	F2_F4	F3_F4
CQ	.2	.4	0.	.6	.6	.6	1.	.6	.6	.6

ture sets in Table XII the letter “F” in front of each feature was omitted.

The highlighted feature sequences F2_F3 and F2_F4 in Table XI and the feature sets 1, 2_3, 4 and 1, 2.4, 3 in Table XII correspond to the highest classification accuracy.

For this example, the feature sequences F2_F3 and F2_F4 produced the best classification accuracy. Among others, feature sequences of size three were considered, however, without success due to the small number (four) of features in the original data set of Table III. Wider data sets (containing more features) offer more opportunity for the definition of feature sequences with multiple original features, thus offering more opportunities for improvement of classification accuracy.

The benefits offered by feature sequences are illustrated with a data set from semiconductor industry. The case study was to accomplish two goals. The first goal was to develop a model associating process parameters with the product quality. Such a model is to be used to predict the quality of products before the actual process takes place. The second goal was to reduce the number of process parameters to be actively controlled based on this model. Prior experimentation with statistical and computational intelligence tools has not produced a satisfactory solution to the two goals.

The industrial data set includes 86 objects (observations), each containing 93 feature values (process parameters). The data set has been collected at random intervals over a year period, therefore, the conditions of the process might have varied

across observations. The data set was essentially complete as all feature values were collected by a computerized data acquisition system. Product quality has been selected as a decision for data analysis. The decision value was the only feature that could not be automatically determined, rather it had to be manually assigned based on the data collected in addition to the 93 features. Determining decision values is a problem in itself that needs a separate consideration.

For the purpose of this study three decision values were selected, $D = \{N \text{ (Negative), } Z \text{ (Zero), } P \text{ (Positive)}\}$. The assignment values to the decision could involve inconsistency between the adjacent classes due to human error, e.g., for a particular product the decision $D = P$ could be assigned rather than $D = N$. However, misrepresentations across two decision classes, e.g., $D = N$ rather than $D = P$ are not possible. The number of quality levels (three) was selected for the purpose of this study and its use in industrial practice.

IV. INDUSTRIAL CASE STUDY

To evaluate the impact of feature sequences on classification accuracy, seven different cases (data sets) have been considered grouped into two computing scenarios.

A. Computing Scenario 1

Case 1: Original (Not Transformed) Features: The set $\{F7, F35, F40\text{--}F44, F47, F55, F56, F68\text{--}F72, F73, F78\}$ of features that were viewed as high priority control parameters has been selected and decision rules were extracted with a rough set algorithm.

The 1-out-of- n ($n = 86$ objects) cross-validation scheme was used to validate the results produced by the rough set algorithm. The results of cross-validation for Case 1 are shown in Tables XIII and XIV.

Table XIV reports the percentage of objects that have been classified correctly, incorrectly, or fall into the “None” category for the three decision values $D = \text{Negative, Zero, or Positive}$.

Case 2: Feature Sequence F68_F72: In this case the data set from Case 1 was transformed by replacing the five individual features F68 through F72 with the feature sequence $\{F68.F69.F70.F71.F72\}$ denoted as F68_72. In total 31 different values of the feature sequence F68_72 have been used in the entire training data set.

The feature sequence F68_72 values are shown at the bottom of the next page.

In fact this data mining study has made the company aware of the significance of the feature sequence F68_72 in process control. The classification accuracy of each individual feature F68 through F72 is zero, while the classification accuracy of the

TABLE XII
SUMMARY OF 1-OUT-OF- n ($n = 5$) VALIDATION RESULTS

Feature set	1, 2, 3, 4	1_2, 3, 4	1_3, 2, 4	1_3, 2, 4	1_4, 2, 3	1_2_3, 4	1_2_4, 3	1_2_3_4	1_2, 2_3, 4
Correct	20%	40%	20%	20%	20%	60%	60%	40%	40%
Incorrect	60%	60%	60%	60%	40%	40%	40%	40%	60%
None	20%	0%	20%	20%	40%	0%	0%	20%	0%

TABLE XIII
ABSOLUTE CLASSIFICATION ACCURACY FOR CASE 1

D	N	Z	P	None
N	6	1	11	0
Z	1	10	16	0
P	5	7	29	0

feature sequence F68_72 is 24.4%, which implies that 24.4% of all objects in the training set can be uniquely identified by this feature.

The cross-validation results for this case are shown in Tables XV and XVI.

The cross-validation results of computing scenario 1 indicate that the rules generated from the transformed data set of Case 2 produced better prediction accuracy than those of Case 1. Of particular interest is the number of correctly predicted decisions $D = \text{Positive}$, which is 32 in Table XV versus 29 in Table XIII. The average percentage of correctly predicted outcomes is also higher for Case 2, as shown in Table XVI, the average $Av = 54.65\%$ versus 52.33% in Table XIV. The third important measure is the total number of correctly predicted outcomes (sum of diagonal elements in Tables XV and XVII), which is 45 ($6 + 10 + 29$) in Case 1 and 47 ($3 + 12 + 32$) in Case 2.

In computing scenario 2, four different cases involving different number of feature sequences were considered. Each case was generated from the same original data set.

B. Computing Scenario 2

Case 3: Original Features: In this case, the data set with the following features was considered {F6, F7–F11, F25, F35, F40–F44, F47, F55, F56, F57, F63–67, F68–F72, F73–F77, F78–F82}. This feature set is a superset of the set of Case 1.

The cross-validation results are shown in Tables XVII and XVIII.

Case 4: Feature Sequence F68_F72: The data set with the following features was considered {F6, F7–F11, F25, F35, F40–F44, F47, F55, F56, F57, F63–F67, F68_F72, F73–F77, F78–F82}.

The feature sequence F68_72 is defined in Case 2 and the cross-validation results are shown in Tables XIX and XX.

The average percentage of correctly classified objects (under "Correct" in Table XX) for Case 4 is not better than that of Case 3 (Table XVIII), however the number of correctly predicted outcomes with $D = N$ is higher, which implies that feature se-

TABLE XIV
CLASSIFICATION ACCURACY FOR CASE 1

	Correct	Incorrect	None
N	33.33%	66.67%	0%
Z	37.04%	62.96%	0%
P	70.73%	29.27%	0%
Av	52.33%	47.67%	0%

TABLE XV
ABSOLUTE CLASSIFICATION ACCURACY FOR CASE 2

D	N	Z	P	None
N	3	2	12	1
Z	1	12	13	1
P	3	3	32	3

TABLE XVI
CLASSIFICATION ACCURACY FOR CASE 2

	Correct	Incorrect	None
N	16.67%	77.78%	5.56%
Z	44.44%	51.85%	3.70%
P	78.05%	14.63%	7.32%
Av	54.65%	39.53%	5.81%

TABLE XVII
ABSOLUTE CLASSIFICATION ACCURACY FOR CASE 3

D	N	Z	P	None
N	6	1	11	1
Z	2	10	15	1
P	1	3	36	1

TABLE XVIII
CLASSIFICATION ACCURACY FOR CASE 3

	Correct	Incorrect	None
N	33.33%	66.67%	0%
Z	37.04%	62.96%	0%
P	87.80%	9.76%	2.44%
Av	60.47%	38.37%	1.16%

TABLE XIX
ABSOLUTE CLASSIFICATION ACCURACY FOR CASE 4

D	N	Z	P	None
N	1	2	10	5
Z	1	12	12	2
P	1	2	34	4

quences may allow for shifting the number of correctly pre-

0_0_0_0_0,0_0_0_0_1,0_0_0_0_3,0_0_0_1_1,0_0_2_0_0,0_0_3_3_3,0_1_0_0_0,0_1_1_0_0,0_2_0_0_0,0_2_0_0_1,0_2_2_0_0,1_0_0_0_0,1_0_0_0_3,1_0_0_1_1,1_1_0_0_0,1_1_0_0_1,1_1_0_0_3,1_1_1_0_0,1_1_1_0_1,1_1_1_0_3,1_1_1_1_1,1_1_1_2_0,2_0_0_0_0,2_0_0_0_1,2_0_2_2_0,2_2_0_0_0,2_2_2_0_0,2_2_2_0_1,2_2_2_2_0,2_2_2_2_1,2_2_3_3_3

TABLE XX
CLASSIFICATION ACCURACY FOR CASE 4

	Correct	Incorrect	None
N	5.56%	66.67%	27.780%
Z	44.44%	48.15%	7.41%
P	82.93%	7.32%	9.76%
Av	54.65%	32.56%	12.79%

TABLE XXI
ABSOLUTE CLASSIFICATION ACCURACY FOR CASE 5

D	N	Z	P	None
N	4	1	13	0
Z	1	12	12	2
P	1	0	37	3

TABLE XXII
CLASSIFICATION ACCURACY FOR CASE 5

	Correct	Incorrect	None
N	22.22%	77.78%	0%
Z	44.44%	48.15%	7.41%
P	90.24%	2.44%	7.32%
Av	61.63%	32.56%	5.81%

dicted outcomes among different decision categories. The average “Incorrect” value for Case 4 (Table XX) is the highest of all four cases considered in computing scenario 2. The feature sequence sets defined in cases 5 through 7 to be presented next resulted in better values of “Correct” than that of Case 3. The latter implies that feature sequences tend to reduce the number of incorrectly classified objects.

Case 5: Feature Sequences F63_F67 and F68_F72: In this case the data set with the following features was considered {F6, F7–F11, F25, F35, F40–F44, F47, F55, F56, F57, F63_F67, F68_F72, F73–F77, F78–F82}.

The feature sequence F63_F67 is the sequence {F63_F64_F65_F66_F67} and its values are as follows: 4.4.4.4.3, 5.4.4.4.3, 5.5.5.5.3, 5.5.5.6.3, 6.6.4.4.3, 6.6.6.6.3.

The cross-validation results are shown in Tables XXI and XXII.

Case 6: Feature Sequences F63_F67, F68_F72, and F73_F77: The data set with the following features was considered {F6, F7–F11, F25, F35, F40–F44, F47, F55, F56, F57, F63_F67, F68_F72, F73_F77, F78–F82}.

The feature sequence {F73_F74_F75_F76_F77} denoted as F73_F77 has the following values:
 - 100.- 100.- 100.- 50.50, -25.- 25.- 50.- 50.50,
 - 70.-70.- 70.- 70.70, -50.- 50.- 50.- 50.50,
 - 50.- 50.- 50.-30.30, - 50.- 50.-50.- 30.20, - 50.-
 -50.- 30.- 30.30,
 -50.- 50.- 30.- 30.20, - 30.- 30.- 30.- 50.50, - 5
 - 5.-30.- 50.50, - 30.-30.- 30.- 30.50,
 -30.- 30.- 30.- 30.30, - 29.- 31.- 30.- 70.70,
 - 20.- 20.-50.- 50.50, - 15.-25.- 40.- 40.40,
 -100.- 100.- 100.- 30.20.

The cross-validation results are included in Tables XXIII and XXIV.

TABLE XXIII
ABSOLUTE CLASSIFICATION ACCURACY FOR CASE 6

D	N	Z	P	None
N	3	0	14	1
Z	1	12	14	0
P	1	1	38	1

TABLE XXIV
CLASSIFICATION ACCURACY FOR CASE 6

	Correct	Incorrect	None
N	16.67%	77.78%	5.56%
Z	44.44%	55.56%	0%
P	92.68%	4.88%	2.44%
Av	61.63%	36.05%	2.33%

TABLE XXV
ABSOLUTE CLASSIFICATION ACCURACY FOR CASE 7

D	N	Z	P	None
N	5	1	11	1
Z	2	11	14	0
P	1	1	37	2

TABLE XXVI
CLASSIFICATION ACCURACY FOR CASE 7

	Correct	Incorrect	None
N	27.28%	66.67%	3.49%
Z	40.74%	59.26%	5.56%
P	90.24%	4.88%	0%
Av	61.63%	34.88%	3.49%

The results in Table XXIII indicate that of all 41 objects in the training data set, the rules generated for Case 6 resulted in correct classification of 38 objects and misclassification of only 3 objects with decisions $D = Z$, $D = N$, and $D = \text{None}$ (no decision). This is a small error in particular that there was some arbitrariness involved in labeling the decisions D at the three levels. The only notable error in Table XXIII is that of classifying an object with $D = P$ as an object with $D = N$, which can be explained by incompleteness of the training set, human error, or possibly inadequacy of features.

Case 7: Feature Sequences F63_F67, F68_F72, F73_F77, and F78_F82: In this case the data set with the following features was considered {F6, F7–F11, F25, F35, F40–F44, F47, F55, F56, F57, F63_F67, F68_F72, F73_F77, F78_F82}.

The feature sequence {F78_F79_F80_F81_F82} = F78_F82 has the following values:
 50.50.30.30.30, 50.50.50.30.20, 30.30.30.50.50,
 70.70.70.70.70,
 50.50.50.50.50, 50.50.50.50.20, 50.50.50.30.30, 50.
 50.30.30.20, 5.5.30.50.50, 30.31.30.70.70,
 30.30.30.30.50, - 30.-30.- 30.- 30.30,
 25.25.50.50.50, 20.20.50.50.50, 15.25.
 40.40.40, 100.100.100.50.50,
 100.100.100.30.30, 100.100.100.30.20.

The cross-validation results are shown in Tables XXV and XXVI.

The summary of cross-validation results for Cases 3–7 of computing scenario 2 are included in Tables XXVII and

TABLE XXVII
SUMMARY OF CORRECTLY PREDICTED OUTCOMES FOR CASES 3–7

	Case 3	Case 4	Case 5	Case 6	Case 7
N	6	1	4	3	5
Z	10	12	12	12	11
P	36	34	37	38	37

TABLE XXVIII
SUMMARY OF AVERAGE CLASSIFICATION ACCURACY
FOR CASES 3–7

	Case 3	Case 4	Case 5	Case 6	Case 7
Correct	60.47%	54.65%	61.63%	61.63%	61.63%
Incorrect	38.37%	32.56%	32.56%	36.05%	34.88%
None	1.16%	12.79%	5.81%	2.33%	3.49%

XXVIII. The best values of attained absolute accuracy numbers are highlighted. The highlighted entries in Table XXVII indicate in Case 3 the maximum number of correctly predicted examples in $D = N$ category has been generated. Case 6 attains two maxima, one for $D = Z$ and the other for $D = P$.

The classification accuracy in Table XXVIII show that the maxima for “Correct” are attained for Cases 5, 6, and 7 and the minima for “Incorrect” are for Cases 4 and 6.

The cross-validation results of computing scenario 2 indicate that the rules generated based on the transformed data sets of Cases 4, 5, 6, and 7 produced better prediction accuracy than those of Case 3 based on at least one of the following three performance measures.

- 1) Maximum absolute number of correctly predicted decisions in one or more of the three categories, $D = N$, $D = Z$, or $D = P$ (see Table XXVII).
- 2) Maximum percentage of correctly classified objects (“Correct” in Table XXVIII).
- 3) Minimum percentage of incorrectly classified objects (“Incorrect” in Table XXVIII).

One should note that the values $D = Z$ and “Incorrect” (the bold entries of Tables XXVII and XXVIII) for Case 3 are the worst of all the five cases. Behaving similarly are the entries $D = N$ and “Correct” for Case 4.

The results reported in Tables XXVII and XXVIII are by no means global optima. Feature sequences that may produce results of better quality are certainly possible. Further research is needed to develop a better understanding of feature sequences and feature bundling in data mining.

The computational results reported in this section indicate that the values of control parameters can be associated with the decision D in the form of decision rules. Of particular interests are the associations that optimize performance measures [27].

The computational results reported in the paper lead to two important conclusions.

- 1) Feature sequences are a viable way of increasing classification accuracy.
- 2) The best rules involving different sets of features can be combined to maximize the overall classification accuracy, e.g., the decision rules of Case 6 maximizing $D = P$ and $D = Z$, and the rules of Case 3 maximizing $D = N$.

V. CONCLUSION

Data mining offers methodologies and tools for discovery of new knowledge for decision making. Classification accuracy of decisions made with the extracted knowledge depends on the properties of the training data set. In most data mining applications raw data is used for rule extraction. In this paper a new transformation method named feature bundling was introduced. This transformation, when applied to a training data set, enhances classification accuracy of the decision rules generated from this set.

The reason for increased classification accuracy with feature bundling might due to the fact that the associations among features and decisions are stronger than those built on feature values. Although bundling is primarily intended for integer, normative and categorical features, it can be extended to features with continuous values, for example, by using regression functions.

One of many feature bundles, a feature sequence, is analogous to a gene sequence in a chromosome. The study of expression and functionality of gene sequences may offer valuable improvements of classification accuracy of data mining approaches.

APPENDIX

DEFINITIONS OF BASIC DATA MINING TERMS

Rule support is the number of objects in the data set that have the property described by the conditions of the rule.

Rule strength is the number of objects in the data set that have the property described by the conditions and the decision of the rule.

Relative rule strength is the percentage of objects in the data set within the same category that have the property described by the conditions and the decision of the rule.

Discrimination level is the percentage ratio of the rule strength and the rule support.

Absolute classification accuracy for a rule set is the number of correctly classified objects from the test set.

Classification accuracy (CA) for a rule set is the number of correctly classified objects from the test set to all objects in the test set.

Classification quality (CQ) of a feature set is the percentage of all objects in the training data set that can be unambiguously associated with the decision values based on the features in this set [14].

REFERENCES

- [1] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning, Neural, and Statistical Classification*. New York: Ellis Horwood, 1994.
- [2] P. Domingos and M. Pazzani, “Beyond independence: Conditions for the optimality of the simple Bayesian classifier,” in *Mach. Learn.: Proc. 13th Int. Conf.*, Los Altos, CA, 1996, pp. 105–112.
- [3] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.
- [4] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
- [5] P. Clark and R. Boswell, “The CN2 induction algorithm,” *Mach. Learn.*, vol. 3, no. 4, pp. 261–283, 1989.
- [6] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Los Altos, CA: Morgan Kaufmann, 1993.

- [7] P. Auer, R. Holte, and W. Maass, "Theory and application of agnostic PAC-learning with small decision trees," in *Proceedings of the 8th European Conference on Machine Learning*, A. Prieditis and S. Russell, Eds. New York: Springer Verlag, 1995.
- [8] J. Friedman, Y. Yun, and R. Kohavi, "Lazy decision trees," in *Proceedings of the Thirteenth National Conference on Artificial Intelligence*. Cambridge, MA: MIT Press, 1996.
- [9] R. Kohavi, "Wrappers for Performance Enhancement and Oblivious Decision Graphs," Ph.D. dissertation, Comp. Sci. Dept., Stanford Univ., Stanford, CA, 1995.
- [10] D. W. Aha, "Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms," *Int. J. Man-Mach. Stud.*, vol. 36, no. 2, pp. 267–287, 1992.
- [11] R. S. Michalski, I. Bratko, and M. Kubat, Eds., *Machine Learning and Data Mining*. New York: Wiley, 1998.
- [12] R. S. Michalski, I. Mozetic, J. Hong, and N. Lavrac, "The multi-purpose incremental learning system AQ15 and its testing application to three medical domains," in *Proceedings of the 5th National Conference on Artificial Intelligence*. Palo Alto, CA: AAAI, 1986, pp. 1041–1045.
- [13] J. W. Grzymala-Busse, "A new version of the rule induction system LERS," *Fund. Inform.*, vol. 31, pp. 27–39, 1997.
- [14] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Boston, MA: Kluwer, 1991.
- [15] J. Stefanowski, "On rough set approaches to induction of decision rules," in *Rough Sets in Knowledge Discovery I: Methodology and Applications*, L. Polkowski and A. Skowron, Eds. New York: Springer-Verlag, 1998, pp. 501–529.
- [16] L. B. Brooker, "Triggered rule discovery in classifier systems," in *Proceedings of the 3rd International Conference on Genetic Algorithms (ICGA89)*, J. D. Schaffer, Ed. San Mateo, CA: Morgan Kaufmann, 1989, pp. 265–274.
- [17] J. Y. Donnat and J. A. Meyer, "A hierarchical classification system implementing a motivationally autonomous animat," in *Proceedings of the Third International Conference on Simulation of Adaptive Behavior (SAB94)*, D. Cliff, P. Husbands, J. A. Meyer, and S. W. Wilson, Eds. Cambridge, MA: MIT Press, 1994, pp. 144–153.
- [18] S. W. Wilson, "Classifier fitness based on accuracy," *Evol. Comput.*, vol. 3, no. 2, pp. 149–175, 1995.
- [19] T. Kovacs, "What should a classifier system learn?," in *Proceedings of the Congress on Evolutionary Computation 2001*. Piscataway, NJ: IEEE Press, 2001, pp. 775–782.
- [20] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Mach. Learn.*, vol. 40, pp. 203–228, 2000.
- [21] V. Cherkassky and F. Mulier, *Learning from Data—Concepts, Theory, and Methods*. New York: Wiley, 1998.
- [22] P. L. Lanzi, W. Stoltzmann, and S. W. Wilson, Eds., *Learning Classifier Systems: From Foundations to Applications*. New York: Springer, 2000.
- [23] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: Univ. Michigan Press, 1975.
- [24] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.
- [25] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Diego, CA: Academic, 2001.
- [26] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Royal Stat. Soc.*, vol. 36, pp. 111–147, 1974.
- [27] A. Kusiak and C. Kurasek, "Data mining analysis of printed-circuit board defects," *IEEE Trans. Robot. Automat.*, vol. 17, pp. 191–196, Apr. 2001.



Andrew Kusiak (M'90) is a Professor of industrial engineering at the University of Iowa, Iowa City. He is interested in theory and applications of computational intelligence, data mining, and optimization in product development, manufacturing, and healthcare. He has published research papers in journals sponsored by AAAI, IEEE, IIE, INFORMS, ESOR, IFIP, IFAC, IPE, ISPE, and SME. He speaks frequently on international meetings, conducts professional seminars, and consults for industrial corporations. He serves on the editorial boards of 16 journals, and edits book series. He is the Editor-in-Chief of the *Journal of Intelligent Manufacturing*.