Data Mining: Medical and Engineering Case Studies

A. Kusiak*, K.H. Kernstine, J.A. Kern***, K.A. McLaughlin**, and T.L. Tseng***
Department of Industrial Engineering, 4132 SC
****Department of Surgery, 1616-B JCP***
*****Department of Internal Medicine, C33-A GH***
The University of Iowa
Iowa City, Iowa 52242 - 1527
http://www.icaen.uiowa.edu/~ankusiak

Abstract

In the paper the problem of predicting outcomes in medical and engineering applications is discussed. The problem is solved with a data mining approach. The first of the two algorithms presented in the paper generates a solution that is validated with the second algorithm. The validated solution constitutes the final decision.

Keywords

Data mining, rough set theory, autonomous diagnosis, decision making, lung cancer, cost estimation.

1. Introduction

The interest in systems for autonomous decisions in medical and engineering applications is growing, as data is becoming more easily available. Though the two areas – medicine and engineering – appear to be remote in terms of the underlying processes, both face many common challenges. One of the problems of interest to both areas is autonomous prediction. In this paper two instances of the prediction problem are considered, the diagnosis problem in medicine and the cost estimation problem in engineering. Each of the two problems is solved with two independent data mining algorithms.

2. Medical Diagnosis Problem

An SPN (solitary pulmonary nodule) is a lung abnormality that may be cancerous or benign. Over 160,000 people in the US only are affected by lung cancer, and over 90% of them die. It is important that SPNs are diagnosed early and accurately. The clinical diagnosis of SPN using information from noninvasive tests is 40–60% accurate. This figure implies that many patients suspected of malignancy have to undergo biopsy that involves considerable risks (including death) and costs to them. The data mining algorithms discussed in the paper significantly reduce patients' risks and diagnosis costs. In a typical SPN disease occurrence scenario, a nodule is detected on a patient's chest radiograph. As this SPN may be either benign or malignant, further testing is required to determine its exact nature. The diagnosis is perceived to depend on many features, such as the SPN diameter, border character, presence of calcification, patient's age, smoking history, and so on (Lillington 1993). Multiple medical disciplines are involved collecting a large volume of clinical data at different times and locations, with varying accuracy and consistency. Therefore, an approach that fuses information from different sources and intelligently processes large volumes of data is needed. The proposed algorithms use features extracted from data sets of different origins.

The research presented in this paper shows that that number of features (results of noninvasive tests, patient's data, etc.) necessary to diagnose an SPN is smaller that used in current medical practice. At the same time the decisions made are 100% accurate.

Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, May 21-23, 2000, pp. 1-7.

3. Engineering Design Problem

Engineering design involves decisions where parameters, actions, components, and so on are selected (Kusiak 1999). This selection is often made based on prior data, information, or knowledge. The class of selection problems in engineering design is illustrated with the problem of cost estimation (referred here as the prediction problem) of printed circuit boards. To date, numerous models and algorithms have been developed for autonomous predictions based on data corresponding to different characteristics (features). The algorithms proposed in this research are designed to make predictions of high accuracy for a large percentage (possibly 100%) of cases being considered. For cases when an accurate decision could not be automatically generated or confirmed other decision-making modalities could be used, e.g., higher-level prediction systems, humans, or neural networks.

The prediction problem in medicine and engineering design is structured into two phases:

- Learning phase, and
- Decision-making phase

In the learning phase a large data set (object – feature matrix) is transformed into a reduced data set, called here a *decision table*. The number of features and objects in the *decision table* is much smaller than in the original data set. Each object (row) in the original data set may be represented in the *decision table* in more different way. In this paper, the decision-making phase will be emphasized over the learning phase.
The *decision table* is used to make predictions when a new case with unknown outcome arrives with Prediction Algorithm 1. This algorithm compares the feature values of a new object with the feature values of objects represented in the *decision table*. If the match is satisfactory, the new object is assigned an outcome equal to the outcome of the matching object(s). The result produced by Prediction Algorithm 1 is verified with Prediction Algorithm 2. Both algorithms are discussed later in this paper.

4. Literature Review

In this paper, the basic formalisms used in the *learning phase* of autonomous prediction are based on the *rough set theory* (Pawlak 1991), specifically, the concept of *feature extraction*.
The *rough set theory* is one of unique theories in data mining. It has found many applications in industry, service organizations, and healthcare (Kowalczyk and Slisser 1997, and Weiss and Indurkhya 1998). A comprehensive comparative analysis of prediction methods in medical applications is included in Kononenko *et al.* (1998) indicates that automatically generated diagnostic rules outperform the diagnostic accuracy of physicians. The authors' claim is supported by a comprehensive review of the literature on four diagnostic topics: localization of primary tumor, prediction of reoccurrence of breast cancer, thyroid diagnosis, and rheumatoid prediction. Berry and Linoff (1997) and Groth (1998) surveyed some engineering and business applications of data mining.

In this paper the data mining approach has been selected over regression analysis and neural networks. There are three fundamental differences between the latter two approaches and the one discussed in this paper. First, both neural networks and regression make decisions for essentially all cases with an error, while the proposed approach makes accurate decisions when it has sufficient amount of 'knowledge'. Second, neural networks and regression models are 'population based' while the feature extraction approach follows the 'individual (data object) based paradigm'. The two 'population based' tools determine features that are common to a population. The *feature extraction* concept of *rough set theory* identifies unique features of an object and sees whether these unique features are shared with other objects. It is obvious that the two paradigms differ and in general the set of features derived by any of the two paradigms is different. Thirdly, each of the two 'population based' methods uses a fixed set of features to arrive at a decision. In the *feature extraction* approach the same set of features may apply to a group of objects.

Most of the data mining literature emphasizes analysis of the data collected (learning phase). In this paper, novel algorithms for predicting decisions for objects with unknown outcomes are presented.

An important issue in autonomous predictions is that of user's confidence in the results generated in the decision-making phase. The data used for prediction may contain errors. What metrics should be used to measure that confidence? When should the user trust the outcome of a prediction algorithm? There are at least two ways of increasing user's confidence in the outcome: (1) incorporating redundancy in the prediction algorithms (Lee and Morey 1994), and (2) using independent methods to verify the same result (Klein 1993). The two ways of robustness enhancement are explored in this paper.

5. Rough Set Theory

Rough set theory is based on the assumption that data and information is associated with every object of the universe of discourse (Pawlak 1991, 1997). Objects described by an equivalent (or similar) set of selected *features* (attributes) are indiscernible. The models and algorithms to be developed in this research will select a subset of all features necessary to characterize the objects.

The basic construct of rough set theory is called a *reduct* (Pawlak 1991). It is defined as a minimal sufficient subset of features RED $\subseteq$ A such that (Shan *et al.* 1995):

(a) Relation R(RED) = R(A), i.e., RED produces the same categorization of objects as the collection A of all features, and

(b) For any e $\in$ RED, R (RED - {e}) $\neq$ R(A), i.e., a reduct is a minimal subset of features with respect to the property (a).

By definition, *reduct* represents an alternative and simplified way of representing a set of objects. It is easy to see that reduct has the same properties as a key defined in relational database theory (with respect to a specific instance of a relation only). In this context reduct can be called an *empirical key*.

The term *reduct* was originally defined for sets rather than objects with features and decisions (a decision table). Reducts of the objects in a decision table have to be computed with the consideration given to the value of the outcome. The original definition of *reduct* considered features only. In this paper each *reduct* is viewed from four perspectives: feature, feature value, object, and rule perspective (Kusiak 2000). To illustrate the two of these perspectives consider the data in Table 1 for five objects, each with four features F1- F4 and prediction D.

Table 1. Five data objects

Object 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

For example, consider reduct (1) for object 2 of Table 1.

1 x x x 2 (1)

The first four elements in this reduct represent the object features and the last value 2 represents the outcome. The value of feature F1 is 1 and the reduct is referred to as a *single-feature reduct*. The features F2 – F4 marked with "x" are not contained in the reduct. As this reduct has been derived from object 2 we will refer to it as an *o-reduct* (object-reduct). The reduct in (1) can be also expressed as the following decision rule

IF the value of feature F1 = 1 THEN the value of output feature O = 2 (2)

and therefore it is called an *r-reduct* (rule-reduct).

The reduct generation algorithm essentially applies the *reduct* definition to each object (Kusiak 2000). To obtain a *reduct*, one input feature at a time is considered and it is determined whether this feature uniquely identifies the corresponding object. A conflict in the outcome of any two objects disqualifies that feature from producing a single-feature reduct.

Pawlak (1991) proposed to enumerate reducts with one or m – 1 features only. For example, the reduct in (1) uniquely determines an outcome based on a single input feature only.

6. Engineering Case Study

6. 1 Data Set Description

The company has collected over 8,000 records (objects) of historical information. Each information object was described with the following six features F1 through F6 and decision D (see Table 2).

Table 2. Original features

F1: Quantity
F2: Heat sink type
F3: Material
F4: Length (x100)
F5: Width (x100)
F6: Layers
D: Cost

The six features were used to manually predict the component cost. The error rate and time to make that prediction were not acceptable. In addition to he original six features, three supplementary features listed in Table 3 will be considered in data mining.

Table 3. Supplementary features

F7: Vendor number
F8: Lead-time
F9: Component type

The values of these three features were not included in the original database, however, they were available in separate data files.

To analyze the data, 250 objects have been selected from the company's database. The objects were essentially randomly chosen, however, we made sure that the additional data for the features listed in Table 3 were available. The values of features F1, F2, F3, and F4 were discrete while features F4 and F5 were expressed with precision of two decimal points. For the convenience of analysis we have multiplied each of the two feature values by 100. The cost was expressed in dollars. The cost values were grouped in 12 arbitrary classes shown in Table 4.

Table 4. Classes of costs (decision D)

Class 1:	$2 – 11
Class 2:	$14 – 30
Class 3:	$31 – 38
Class 4:	$39 – 43
Class 5:	$44 – 55
Class 6:	$56 – 86
Class 7:	$93 – 150
Class 8:	$159 – 300
Class 9:	$302 – 614
Class 10:	$712 – 970
Class 11:	$1000 – 1295
Class 12:	$1296 – 1925

The cost (outcome) values were grouped in classes for the following reasons:

- Ease of data presentation and analysis
- Company's interest in a price estimate
- Limitations of the prototype software used in some steps of the data analysis
- Limited number of objects (250) considered in the study

The classification quality of each feature in the original 250-object data set was as follows: $C_{F1} = 2.4\%$, $C_{F2} = 29.6$, $C_{F3} = 0\%$, $C_{F4} = 0\%$, $C_{F5} = 70.8\%$, and $C_{F6} = 64.4\%$.

7. Computational Results

7.1 Original engineering data set with six features F1–F6

Reducts: {F1, F5, F6} and {F1, F4, F6}
Classification quality of all features: 96.8%
Classification quality of selected feature sets: {F1, F2, F6} = 64.4%, {F1, F2, F3, F6} = 64.4%, {F1, F2, F3, F4, F6} = 96.8%, {F1, F2, F3, F4, F5, F6} = 96.8%, {F1, F3, F6} = 57.6%, {F1, F3, F4, F6} = 96.8%
Total number of decision rules generated: 131 (127 decision rules and 4 conflicting rules.
Examples of decision rules are presented in Table 5 and 6.

Table 5. Examples of decision rules

```
Decision rule 1. IF (A4 = 385) THEN (D=1); [3, 12.50%, 100.00%]
Decision rule 2. IF (A2 = 1) AND (A1 = 5) THEN (D=1); [4, 16.67%,
100.00%]
Decision rule 3. IF (A4 = 193) THEN (D=1); [4, 16.67%, 100.00%]
Decision rule 4. IF (A5 = 348) THEN (D=1); [1, 4.17%, 100.00%]
Decision rule 5. IF (A4 = 258) THEN (D=1); [2, 8.33%, 100.00%]
```

Table 6. Examples of conflicting rules

```
Decision rule 128. IF (A1 = 25) AND (A4 = 530) THEN (D=2) OR (D=3); [2,
100.00%, 100.00%]
Decision rule 129. IF (A5 = 295) AND (A1 = 1) THEN (D=3) OR (D=4); [2,
100.00%, 100.00%]
```

The reason for getting conflicting rules and ultimately and conflicting predictions with the algorithms is due insufficient number of features describing the data. Note that objects 1, 2, and 4 in Table 7 have identical features, however, dramatically different decisions D.

Table 7. Sample data set

Object No.	F1	F2	F3	F4	F5	F6	D
1	1	1	1	193	213	2	879
2	3	1	1	193	213	2	197
3	1	1	1	1.93	213	2	1,200
4	1	1	1	1.93	213	2	571

Each of the three objects 1, 2, and 4 belongs to a different class (10, 8, and 9) listed in Table 4. It is clearly evident that accurate description of the learning set with the six features is not possible. Incorporating additional features will result is a better description of the learning set as well as more accurate predictions to made in the decision-making phase.

7.2 Expanded data set with six original features F1–F6 and part type (F9)

Reducts: {F1, F2, F5}, {F1, F2, F3, F4, F5}, {F1, F2, F3, F5, F6}, {F9}
Classification quality of all features: 100%
Classification quality of selected feature sets: {F1, F2, F3}= .600, {F1, F2, F9}= 1.000, {F1, F2, F3, F4}= 61.2%, {F1, F2, F3, F4, F6} = 98.4%
Total number of decision rules generated: 82 decision rules (see Table 8). In this case, no conflicting rules were extracted.

Table 8. Examples of decision rules

```
Decision rule 1. (A9 = 69) THEN (D=1); [1, 100.00%, 100.00%]
Decision rule 2. (A9 = 65) THEN (D=2); [5, 8.20%, 100.00%]
Decision rule 3. (A9 = 92) THEN (D=2); [5, 8.20%, 100.00%]
Decision rule 4. (A9 = 47) THEN (D=2); [6, 9.84%, 100.00%]
Decision rule 5. (A9 = 49) THEN (D=2); [3, 4.92%, 100.00%]
```

The decision rules generated in the learning phase are used to make accurate decisions with two prediction algorithms in the decision-making phase. The outcome produced by Prediction Algorithm 1 is verified by Prediction Algorithm 2.

8. Medical Case Study

In the medical case study, we considered data for 50 patients (objects) with known SPN diagnosis. For each patient 18 feature values (mostly clinical test results) were collected. The features considered for the 50 patients could be grouped into two main categories:

- Patients information (e.g., age)
- Test results (e.g., computed tomography)

The two algorithms used in the decision-making phase are outlined next.

8.1 Prediction Algorithm 1

Step 1. Match the object's feature values with the feature values of the decision rules. If the match is satisfactory, go to Step 2.
Step 2. Assign the cost of the object equal to the cost associated with the action of the matching decision rule.
Step 3. Output "No decision has been assigned - More features are needed"

8.2 Prediction Algorithm 2

Step 0. Select a proper feature set.
Step 1. Cluster objects in groups with equal outcomes.
Step 2. Compute distance d_{ij} between a new object and every object in each group.
Step 3. For each group, compute the average distance of the distances d_{ij} of Step 2.
Step 4. Assign the new object a decision corresponding to the group with minimum average distance.

9. Computational Results

The two algorithms were tested on medical and engineering data sets. For the engineering set, 10 objects of unknown outcomes and different number of features were selected. The first set of data included two features only, F1 and F6. Prediction Algorithm 1 has produced three decisions for objects 1, 4, and 10. Two of these outcomes for object 1 and 10 have been confirmed by Prediction Algorithm 2. Despite the fact that only two features were used, the results are quite encouraging. The results generated by Prediction Algorithm 2 had variability ranging from 0 to 8 classes.

Increasing the number of features in the testing set to six has resulted in eight correct predictions by Algorithm 1 and six confirmations by Algorithm 2 as shown in Table 9.

Table 9. Computational results for the engineering data set with six features F1– F6

	D = 1	D = 2	D = 3	D = 4	D = 5	D = 6	D = 7	D = 8	D = 9	D = 10	D = 11	D = 12	Min {(D = 1) - (D = 12)}
Object 1 (D=2)	623.9	408.9	398.1	394.5	509.2	416.3	589.8	516.8	614.2	446.1	478	501.7	4
Object 2 (D=2)	690.9	426.6	431.7	422	480	341.7	513.77	440.8	538.2	370.1	402	425.7	6
Object 3 (D=6)	715.9	448.4	450.8	441.5	487.4	331.5	489.2	415.8	513.2	345.1	377	400.7	6
Object 4 (D=1)	447.8	450.4	579.4	601	720.9	637.6	781.4	745.2	810.2	656	613.3	732.6	1
Object 5 (D=1)	459.4	496.9	627.4	649	768.9	685.6	829.4	793.2	858.2	704	661.3	780.7	1
Object 6 (D=1)	481.6	360.6	501.4	495.6	603	502.6	646.4	610.2	675.21	521	478.3	597.7	2
Object 7 (D=2)	544.8	369.9	528.5	513.3	569.4	430	573.4	537.2	602.2	448	405.3	524.7	2
Object 8 (D=2)	562.4	382.1	541	528.9	571.3	415.6	555.4	519.2	584.2	430	387.3	506.7	2
Object 9 (D=2)	754.5	577.2	595.6	556.1	620.4	681.7	689.8	825.2	832	835.5	666	605.7	4
Object 10 (D=8)	867.6	510.1	376.8	336.6	374.9	292.1	395.1	398	469.1	382.2	368.3	237	6

The shaded cells in the last column of Table 9 indicate the solutions validated by Algorithm 2. The shaded numbers indicate the correctly predicted solutions.

The test file with nine features F1 – F9 has produced 100% correct predictions by the two algorithms.

The results for the medical set are even better. The prediction accuracy was 100% for 91.3% cases considered as illustrated in Table 10.

Table 10. Computational results for the medical data set with 18 features

Test Set	Feature Set	A	B	C
10 internal patients	Classification accuracy	100%	100%	100%
	Diagnostic accuracy	100%	100%	100%
13 additional patients	Classification accuracy	100%	84.6%	92.3%
	Diagnostic accuracy	100%	100%	100%
Total test set (10+13)	Classification accuracy	100%	91.3%	95.3%
	Diagnostic accuracy	100%	100%	100%
Total test set (10+13)	**Classification accuracy**		**91.3%**	
	Diagnostic accuracy		**100%**	

In this case the number of features used to make accurate decisions was less than included in the original data set.

10. Conclusion

The class of prediction problems in medicine and engineering was considered. The two algorithms proposed in this paper are able to make high accuracy predictions for a large percentage of cases tested. The algorithms were tested on medical and industrial data sets.

References

Berry, M.J.A. and Linoff, G. (1997), *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley, New York.

Groth, R. (1998), *Data Mining: A Hands-On Approach for Business Professionals*, Prentice Hall, Upper Saddle River, N.J.

Grzymala-Busse, J.W. (1997), A new version of the rule induction system LERS, *Fundamenta Informaticae*, Vol. 31, pp. 27-39.

Klein, G.A. (1993), A recognition-primed decision (RPD) model of rapid decision making, in G.A. Klein et al. (Eds), *Decision Making in Action: Models and Methods*, Ablex, Norwood, N.J., 138-147.

Kononenko, I. Bratko, I., and Kokar, M. (1998), Application of machine learning to medical diagnosis, in Michalski, RS, Bratko, I and Kubat M. (Eds), *Machine Learning in Data Mining: Methods and Applications*, Wiley, New York, pp. 389-428.

Kowalczyk, W. and Slisser, F. (1997), Modeling Customer Retention with Rough Data Models, *Proceedings of the First European Symposium on PKDD '97,* Trondheim, Norway, pp. 4-13.

Kusiak, A. (1999), *Engineering Design: Products, Processes, and Systems*, Academic Press, San Diego, CA.

Kusiak, A. (2000), *Computational Intelligence in Design and Manufacturing*, John Wiley, New York.

Lee, J.D. and Moray, N. (1994), Trust, self-confidence, and operators' adaptation to automation, *International Journal of Human-Computer Studies*, Vol. 40, pp.153-184.

Lillington G.A. (1993), Management of the Solitary Pulmonary Nodule, *Hospital Practice*, May, 41-48.

Pawlak, Z. (1991), *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, Boston, MA.

Pawlak, Z. (1997), Rough Sets and Data Mining, *Proceedings of the Australiasia-Pacific Forum on Intelligent Processing and Manufacturing of Materials*, Edited by T. Chandra, S.R. Leclair, J.A. Meech, B. Varma, M. Smith, and B. Balachandran, Vol. 1, Gold Coast, Australia, pp. 663-667.

Shan, N., Ziarko,W., Hamilton H.J., and Cercone, N. (1995), Using Rough Sets as Tools for Knowledge Discovery, *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Edited by Fayyad, U.M. and Uthurusamy, R., AAAI Press, Menlo Park, CA, 263-268.

Weiss, SM and Indurkhya, N (1998), *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann, San Francisco, CA.