Inducing and Evaluating Rule-Based Classification Models from Data

Taras Zavaliy

Information Systems and Networks Department, Lviv Polytechnic National University, S. Bandery Str., 12, Lviv, 79013, UKRAINE, E-mail: taras.zavaliy@gmail.com

The psychological testing results, which were analyzed by the author, are used for decision-making about admitting the power engineering specialists to work. The author describes the rough sets approach and its application for mining rules from data tables. These rules form a classification models that are used to classify new examples. The role of hitting fraction in model induction is considered. Success rate and ROC curves are used for model evaluation. All of the analysis was done with free Rosetta software.

Key words – data mining, rule extraction, classification, soft computing, rough sets, ROC curve.

I. Introduction

Methods of machine learning and data mining are commonly used for finding regularities in data, building prediction models, solving classification or clasterization tasks. Data mining itself is a model-building process. Most data mining methods are based on techniques from machine learning, pattern recognition, and statistics. One of the most widely applied in data mining are soft computing methodologies, including fuzzy sets, rough sets, genetic algorithms and neural networks. The use and evolution of these methods is an active research issue [1, 2]. Thus, rough sets approach for data mining is based on rough set theory introduced by Zdzislaw Pawlak [3, 4]. It is used for reducing the amount of features needed for successful model construction. Rough sets methods naturally deal with noise and redundancy in data.

The real-life problem considered by the author is rooted in humanitarian field. The data set containing psychological testing results was collected in 2006. The testing was conducted at one of the Ukrainian power engineering company in order to improve safety and quality of work. The empirical data collected by psychologists was complemented with somewhat biased information from personnel management. The problem of objectivity and quality of personnel monitoring aroused. The decision-making process in this case was also of interest. The rough sets approach was used for data mining and inducing several classification models. By comparing them with each other it was shown that biased information in the data set is affecting regularities in data and can not be used for decision-making.

II. Rough sets

The rough sets approach has proved to be an efficient mathematical tool for managing uncertainty, noise and redundancy of data in a variety of knowledge discovery tasks. A basic principle of a rough set-based learning is to discover redundancies and dependencies between the features of a problem represented in the form of *decision table* – data table with the decision attribute.

The core of the process is finding the minimal subset of attributes called *reduct* or *minimal hitting set*. Attributes from the reduct preserve all original dependencies and discernibility of examples in the decision table. One of the methods for constructing reduct is Boolean reasoning. This method simplifies Boolean discernibility function $g_A(U)$ [5] constructed from data.

One of the algorithms, which implements Boolean reasoning process for building reduct, is Johnson algorithm [6]. It searches for the minimal hitting set and allows approximate solutions being build. Computing approximate reduct is achieved by aborting the algorithm loop when "enough" amount of strongest attributes have been selected to the reduct. *Hitting fraction* (*HF*) is a tool for controlling approximation. Hitting fraction denotes a fraction of subsets s_i of attributes, which where used in construction of the reduct. Such approximate reduct is stronger than the ordinary, because it reflects more general dependencies in data. Using hitting fraction is another mechanism for getting rid of noise and imprecision in data.

Next step after finding reduct of data table is to generate decision rules and calculate some statistics about them. These rules are used in classification task.

III. Classification task

A classification model consists of a set of rules *RUL* extracted from data and a set *P* of parameters assigned to them. We can also use parameter *HF* when computing approximate reducts. In the case of binary classification a threshold τ for classifying example to a particular class is defined. So, classification model is

$$M = \langle RUL, P, HF, \tau \rangle. \qquad (1)$$

Several numerical parameters can be associated with a decision rule $\alpha \rightarrow \beta$ (IF α THEN β). These parameters are used in classification process directly to rank rules that correspond to the new example.

1. *Support*(α→β) of the rule shows the number of examples in the training data set that have both properties α and β.

2. *Accuracy*(α→β) measures fraction of the rules matching α while having the same conclusion β,

$$\text{accuracy}(\alpha \rightarrow \beta) = \frac{\text{support}(\alpha \rightarrow \beta)}{\text{support}(\alpha)}. \quad (2)$$

3. *Coverage*(α→β) shows how large is the support basis of class β defined by the rule,

$$\text{coverage}(\alpha \rightarrow \beta) = \frac{\text{support}(\alpha \rightarrow \beta)}{\text{support}(\beta)}. \quad (3)$$

These parameters not only describe the quality of the rules, but impact the classification outcome. Classification is usually implemented in the form of voting process [5]. The results of classifying test examples are presented with confusion matrix (CM), in which *True*(X_i) at the main diagonal denotes the number of examples correctly recognized as be-

longing to some class X_i, $False(X_i)$ denotes the number of incorrect classifications to the class X_i.

When applying induced model to the classification of test examples we obtain a *success rate* (*SR*) of correct predictions. The success rate for the multiple-class classification task is defined as (4).

$$SR = \frac{\sum_i True(X_i)}{\sum_i True(X_i) + \sum_i False(X_i)} \qquad (4)$$

In the two-class classification process one of the class, say X_1, casts some normalized number of votes. If this number is greater than predefined threshold τ, than example gets classified to X_1, otherwise it belongs to X_0. A *receiver operating characteristic* (ROC) curve is a graphical representation of how good the classifier separates examples in decision class X_1 from examples in decision class X_0. An ROC curve captures the behavior of classifier as the threshold τ is varied across the full spectrum of possible values. For each value of the threshold τ we obtain a different 2×2 confusion matrix when we classify test examples.

To plot a ROC curve, true positive rate and false positive rate are calculated for each τ. True positive rate represents fraction of test examples correctly predicted as belonging to X_1 class.

$$TPR(\tau) = \frac{True(X_1)}{True(X_1) + False(X_0)} \qquad (5)$$

False positive rate represents fraction of examples incorrectly predicted to belong to X_0 class.

$$FPR(\tau) = \frac{False(X_1)}{False(X_1) + True(X_0)} \qquad (6)$$

The area under the ROC curve (*AUC*) is a measure of how well the classifier is able to discern examples in X_1 from examples in X_0.

$$AUC = \int_0^1 TPR(\tau) dFPR(\tau) \qquad (7)$$

An *AUC* of 0.5 shows that classifier has no ability to discern classes, while an *AUC* of 1 represents perfect discrimination.

IV. Objectives

The experimental setup consisted of data table with psychological testing results. The main goal of experiments was to show if subjective evaluation was affecting the decision process in the company. This was achieved through building and comparing several classification models using rough sets approach implemented in the Rosetta software system [6]. The model induction process involved model approximation to deal with low quality of the data and achieve acceptable results. Approximate models were induced with hitting fraction set to be less than 1, and precise model had hitting fraction *HF*=1. Comparing different models using ROC curves, *AUC* and success rates revealed the optimal solution and led to some assumtions about quality of the data. Conclusions about usefulness of feature reduction were also made.

V. Experiments

The decision table with psychological testing results contained 188 examples and 38 attributes. Conditional attributes contained symbolic information such as employee's job title, job place, age, experience, as well as specific testing results – reaction speed, memory capacity, concentration ability, skills, motivations etc.

Three decision attributes were defined – "successfulness", "reliability" and "aptitude". The table had no missing values. Six models were induced for decision attributes "reliability" and "aptitude" using three different approximation levels. Training and test data sets contained 94 different examples each. Models #1−3 were build for decision attribute "aptitude", and models #4−6 were built to predict "reliability" attribute. The results of model building and evaluation are summarized in Table 3. Sample rules generated from the data and having *coverage*(α)>10% are listed below.

1. age_group(1) ∧ integrated_score(good) → aptitude(good);
2. age_group(2) ∧ integrated_score(good) ∧ ambitions(average) → aptitude(good);
3. job_place(PES) ∧ attention(average) → reliability(good).

Table 1 shows the confusion matrix with the numbers of correct and incorrect predictions of classes made by precise model #1. This model failed to classify 22 examples out of total 94 and marked them as "undefined".

TABLE 1

CONFUSION MATRIX FOR PRECISE MODEL #1

		Predicted class			
		excellent	*good*	*average*	*poor*
Actual class	*excellent*	2	2	0	0
	good	3	49	2	0
	average	1	7	5	0
	poor	0	1	0	0

Table 2 represents confusion matrix for approximate model #3. This model failed to classify only one example from the test set of 94 examples.

TABLE 2

CONFUSION MATRIX FOR APPROXIMATE MODEL #3

		Predicted class			
		excellent	*good*	*average*	*poor*
Actual class	*excellent*	1	4	0	0
	good	3	66	1	0
	average	0	6	10	1
	poor	0	0	1	0

TABLE 3

PARAMETERS OF CONSTRUCTED MODELS

Model	Hitting fraction	Reduct length	Rules count	*SR*	*AUC*
Model #1	1.0	4	60	0.6	0.68
Model #2	0.99	3	32	0.78	-
Model #3	0.96	2	13	0.82	0.81
Model #4	1.0	5	86	0.05	-
Model #5	0.98	3	45	0.35	-
Model #6	0.91	2	16	0.5	0.54

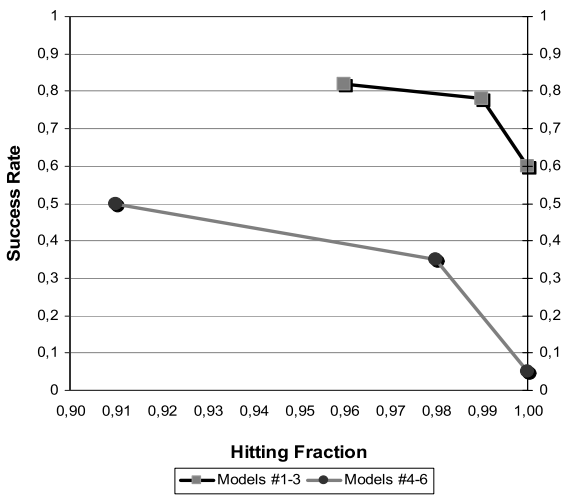

Fig. 1. Performance of classification models depending on approximation level (*HF*).

Fig. 1 represents the performance of six classifiers when classifying 94 test examples.

The next step was to verify classification models using ROC analysis. Figure 2 shows ROC curves plotted for three models – precise model #1, approximate model #3 and approximate model #6.

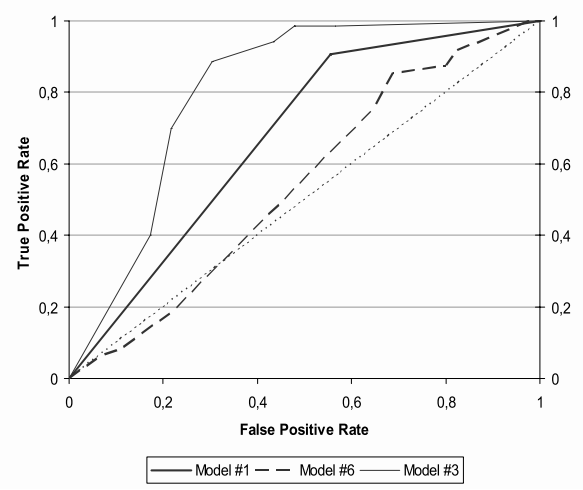

Fig. 2. ROC curves representing ability of the constructed models to recognize class X_1=*good*.

It is obvious that approximate model #3 performs better in recognizing examples from class X_1=*good*. The curve for approximate model dominates the rival curve for #1 model. Comparing the *AUC* values also shows superiority of the approximate model − 0.81 against 0.68 (see Table 3). Meanwhile, model #6 built to predict "reliability" attribute shows very low performance (*AUC*=0.54) in recognizing examples from X_1=*good* class. This can be explained by the fact that "reliability" and "successfulness" attributes in the data set were defined by the personnel management, not by psychologists. Therefore, these attributes are "disconnected" from the rest of the data and are not involved in meaningful regularities. These results obviously reveal the subjective bias in the data table.

Conclusions

The outcome of the research shows that using rough sets approach for machine learning tasks may lead to very useful results. Low quality of training data or insufficient amount of training examples may be compensated by computing approximate solutions. Rough sets methods allow significantly reducing feature set, enabling us to extract more general rules from data. These rules form classification models which show good results classifying test examples.

The data used for rule extraction experiments were not of the highest quality and 94 training examples were not sufficient for high-quality model induction. Yet it was possible to achieve 82% success rate for one of the constructed models. All of the classification models were compared using ROC curves, *AUC* value and success rate. It will be useful applying cross-validation in future experiments for more unbiased results. Using these results for decision-making in future psychological testing of workers and personnel monitoring is of great interest.

References

[1] Mitra S., Pal S. K., Mitra P., "Data mining in soft computing framework: a survey", *IEEE Transactions on Neural Networks,* Vol. 13, Issue 1, 2002.

[2] Wang G., Liu Q., Yao Y., Skowron A. (Eds.), "Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing", *Proceedings of 9th International Conference, RSFDGrC-2003*, Springer, 2003.

[3] Komorowski J., Polkowski L., Skowron A., "Rough Sets: A Tutorial", *// Eds. S. K. Pal and A. Skowron, Rough Fuzzy Hybridization: A New Trend in Decision-Making*, Spriner-Verlag, Singapore, 1998.

[4] Polkowski L., "Rough Sets: Mathematical Foundations", Physica-Verlag, Heidelberg, NY, 2002.

[5] Øhrn A., "Discernibility and Rough Sets in Medicine: Tools and Applications", *PhD thesis*, Norwegian Univ. of Science and Technology, Dep. of Computer and Information Science, 1999.

[6] Øhrn A., "ROSETTA Technical Reference Manual", 2001, http://www.lcb.uu.se/tools/rosetta/.