Noise Models in Inductive Concept Formation

Marina V. Fomina, Alexander P. Eremeev and Vadim V.Vagin
Moscow Pover Engeneering Institute, Technical University, Krasnokazarmennaja Street, Moscov, Russia

Keywords: Knowledge Acquisition, Information Generalization, Data Mining, Knowledge Discovery, Decision Tree, Noise Model, Learning Sample.

Abstract: The problem of information generalization with account for the necessity of processing the information stored in real data arrays which may contain noise is considered. Noise models are presented, and a noise effect on the operation of generalization algorithms using the methods of building decision trees and forming production rules is developed. The results of program modeling are brought about.

1 INTRODUCTION

Knowledge discovery in databases (DB) is important for many technical, social, and economic problems. Up-to-date DBs contain such a huge quantity of information that it is practically impossible to analyze this information manually to acquire valuable knowledge for decisions making. The systems for automatic knowledge discovery in DB are capable of analyzing "raw" data and presenting the extracted information faster and with more success than an analyst could find it himself. At first we consider setting up the generalization problem and the methods of its decision. Then the noise models and prediction of unknown values in accordance with the nearest neighbour method in learning samples are viewed. And finally modeling the algorithm of decision tree with the combination of forming production rules in the presence of noise and results of program simulation are given.

2 GENERALIZATION PROBLEM

Knowledge discovery in DB is closely connected with the solution of the inductive concept formation problem or the generalization problem.
Let us give the formulation of feature-based concept generalization (Vagin et al., 2008).

Let $O = \{o_1, o_2, ..., o_n\}$ be a set of objects that can be represented in an intelligent decision support system (IDSS). Each object is characterized by r attributes: A_1, A_2, ... , A_r. Denote by $Dom(A_1)$, $Dom(A_2)$, ... , $Dom(A_r)$ the sets of admissible values of features where $Dom(A_k)=\{x_1,\ x_2,\ ...\ x_{q_k}\ \}$, $1 \le k \le r$, and q_k is the number of different values of the feature A_k. Thus, each object $o_i \in O$, $1 \le i \le n$ is represented as a set of feature values, i.e., $o_i = \{x_{i1}, x_{i2}, \dots, x_{ir}\}$, where $x_{ij} \in Dom(A_k)$, $1 \le j \le q_k$. Such a description of an object is called a feature description. Quantitative, qualitative, or scaled features can be used as object features (Vagin and Eremeev, 2009).

Let O be the set of all objects represented in a certain IDSS; let V be the set of positive objects related to some concept and let W be the set of negative objects. We will consider the case where $O = V \cup W$, $V \cap W = \varnothing$. Let K be a non-empty set of objects such that $K = K^+ \cup K^-$, where $K^+ \subset V$ and $K^- \subset W$. We call K a learning sample. Based on the learning sample, it is necessary to build a rule separating positive and negative objects of a learning sample.

Thus, the concept is formed if one manages to build a decision rule which, for any example from a learning sample, indicates whether this example belongs to the concept or not. The algorithms that we study form a decision in the form of rules of the type "IF condition, THEN the desired concept." The condition is represented in the form of a logical function in which the boolean variables reflecting the feature values are connected by logical connectives. Further, instead of the notion "feature" we will use the notion "attribute". The decision rule is correct if, in further operation, it successfully recognizes the objects which originally did not

V. Fomina M., P. Eremeev A. and V. Vagin V..
Noise Models in Inductive Concept Formation.
DOI: 10.5220/0004420104440450
In *Proceedings of the 15th International Conference on Enterprise Information Systems* (ICEIS-2013), pages 444-450
ISBN: 978-989-8565-59-4

belong to the learning sample.

The presence of noise in data changes the above setting up of the generalization problem both at the stage of building decision rules and at the stage of the object classification. First of all, the original learning sample *K* is replaced by the sample *K′* in which distorted values or missing values of features occur with a certain probability. We consider the solution of the concept generalization problem using the methods of decision trees (Quinlan, 1986, 1996) including binary decision trees (Breiman et al., 1984).

3 GENERALIZATION ALGORITHMS BASED ON DECISION TREES

The purpose of the given paper is to study a noise influence on the work of generalization algorithms that build decision trees.

The decision tree T is a tree in which each non-final node accomplishes checking of some condition, and in case a node is finite, it gives out a decision for the element being considered. In order to perform the classification of the given example, it is necessary to start with the root node. Then, we go along the decision tree from the root to the leaves until the final node (or a leaf) is reached. In each non-final node one of the conditions is verified. Depending on the result of verification, the corresponding branch is chosen for further movement along the tree. The solution is obtained if we reach a final node. Decision tree may be transformed into a set of production rules.

Research of noise effect on the operation of generalization algorithms has been performed on the basis of comparative analyses of two known algorithms C 4.5 and CART.

The algorithm C4.5 as its predecessor ID3 suggested by J.R.Quinlan (Quinlan, 1986, 1996) refers to an algorithm type building the classifying rules in the form of decision trees. However, C4.5 works better than ID3 and has a number of advantages:

- numerical (continuous) attributes are introduced;
- nominal (discrete) values of a single attribute may be grouped to perform more effective checking;
- subsequent shortening after inductive tree building based on using a test set for increasing a classification accuracy.

The algorithm C 4.5 is based on the following recursive procedure:

An attribute for the root edge of a tree T is selected, and branches for each possible values of this attribute are formed.

The tree is used for classification of learning set examples. If all examples of some leaf belong to the same class, then this leaf is marked by a name of this class.

If all leafs are marked by class names, the algorithm ends. Otherwise, an edge is marked by a name of a next attribute, and branches for each of possible values of these attribute are created, go to step 2.

The criterion for choosing a next attribute is the gain ratio based on the concept of entropy (Quinlan, 1996).

In the algorithm CART (Breiman et al., 1984), building a binary decision tree is performed. Each node of such decision tree has two descendant. At each step of building a tree, the rule that shares a set of examples from a learning sample into two subsets is assigned to a current node. In the first subset, examples are entered where a rule is performed, and the second subset includes examples where a rule does not performed. Accordingly for the current node, two descendant nodes are formed and the procedure is recursively repeated until a tree will be obtained. In this tree the examples of a single class are assigned to each final node (tree leaf).

The most difficult problem of the algorithm CART is a selection of best checking rules in tree nodes. To choose the optimal rule, there is used the assessment function of partition quality for a learning set introduced in (Breiman et al., 1984).

The important distinction of the algorithm CART from other algorithms of building the decision trees is the use the mechanism of tree cutting. The cutting procedure is necessary to obtain the tree of an optimal size with a small probability of erroneous classification.

4 NOISE MODELS

Assume that examples in a learning sample contain noise, i.e., attribute values may be missed or distorted. Noise arises due to following causes: incorrect measurement of the input parameters; wrong description of parameter values by an expert; the use of damaged measurement devices; and data lost in transmitting and storing the information (Mookerjee et al., 1995). Our purpose is to study noise effect on the functioning C 4.5 and CART algorithms.

One of basic parameters of research is a noise level. Let a learning sample K ($|K| = m$) be represented in the table with m rows and r columns, such table has $N=m\cdot r$ of cells. Each table row corresponds to one example and each column – to certain informative attribute. A noise level is a magnitude p_0, showing that an attribute value in a learning or test set will be distorted. So, among all N cells, $N\cdot p_0$ of cells at the average will be distorted. Modeling a noise includes noise models and ways of their entering as well.

For research, two noise models were chosen: "absent values" and "distorted ones". In the first case for the given noise level with probability p_0, a known attribute value is removed from a table. The second variant of entering a noise is linked with substitution of a known attribute value for another one that may be wrong for the given example. Values for replacement are chosen from domains $Dom(A_k)$, $1 \le k \le r$, where p_0 sets up a probability of such substitution.

At entering a noise of the type "absent values", it is necessary also to select a way of treating absent values. In the paper two ways are considered: omission of such example and restoring absent values on the "nearest neighbours" method (Vagin and Fomina, 2011).

There are several ways of entering a noise in learning sets (Quinlan, 1986). Let us consider three ways of entering a noise into a table.

1. Noise is entered evenly in the whole table with the same noise level for all attributes.
2. Noise of the given level is entered evenly in one or several explicitly indicated attributes. Entering a noise into the single table column, the content of which is the most important attribute (root node), is an extreme case here.
3. The new way of irregular noise entering in a table was offered. Here a noise level for each column (informative attribute) depends on a probability of passing an accidentally selected example through a tree node marked by this attribute.

   We have:

- a sum noise entered into a table corresponds to the given noise level;
- all informative attributes, values of which are checked in nodes of a decision tree, are put on distortions;
- the more "important" an attribute the higher a distortion level of its values.

Principles of noise level account for the third irregular model are proposed. Let the decision tree T have been built on the basis of the learning sample K. Evidently, an accidentally selected example will passes far from through all nodes. Hence, our problem is to efficiently distribute this noise between table columns (attributes) in correspondence with statistical analysis of DBs having a given average noise level p_0.

For each attribute A_k, find a factor of the noise distribution S_k according to a probability of passing some example through the node marked A_k. Clearly, each selected example from K will pass through the root node of a decision tree. Therefore the value 1 is assigned to the factor S_k of the root attribute.

All other tree nodes which are not leafs have one ancestor and some descendants. Let one such node be marked by attribute A_i and have the ancestor marked by A_q. The edge between that nodes is marked by the attribute value x_j where $x_i \in Dom(A_q)$. Let m be the example quantity in K and m_j be the example quantity in K satisfying to the condition: attribute value for A_q is equal to x_j.

Then the factor of noise distribution

$$S_{A_i} = S_{A_q} \frac{m_j}{m}.$$

The value 0 is assigned to all factors for attributes not using in a decision tree. Introduce the norm

$$S = \sum_{i=1}^{r} S_{A_i}$$

Thus, each attribute A_i, will be undergone to influence of a noise where a noise level is

$$d_{A_i} = \frac{S_{A_i}}{S} \cdot p_0 \cdot r$$

Here p_0 is a given noise level, r is an attribute quantity.

It is easy to see that $\left(\sum d_{A_i}\right)/r = p_0$, i. e. the average noise level is the same as the given one.

Further, we consider the work of the generalization algorithm in the presence of noise in original data. Our purpose is to assess the classification accuracy of examples in a test sample by increasing a noise level in this sample.

5 RESTORING ABSENT VALUES IN A LEARNING SAMPLE

Let a sample with noise, K', be given; moreover, let the attributes taking both discrete and continuous values be subject to distortions. Consider the problem of using the objects of a learning sample K'

in building a decision tree *T* and in conducting the examination using this tree.

Let $o_i \in K'$ be an object of the sample;

$o_i = <x_{i1}, \ldots, x_{ir}>$. Among all the values of its attributes, there are attributes with the value *Not known*. These attributes may be both discrete and continuous.

Building a decision tree while having examples with absent values leads to multivariant decisions. Therefore, we try to find the possibility for restoring these absent values. One of the simplest approaches is a replacement of an unknown attribute value on the average (mean or the mode, *MORM*). Another possible approach is the nearest neighbours method which was proposed for the classification of the unknown object *o* on the basis of consideration of several objects with the known classification nearest to it. The decision on the assignment of the object *o* to one or another class is made by information analysis on whether these nearest neighbours belong to one or another class. We can use even a simple count of votes to do this. The given method is implemented in the algorithm of restoring RECOVERY that was considered in detail in (Vagin and Fomina, 2010), (Vagin et al., 2008). Since the solution in this method explicitly depends on the object quantity, we shall call this method as the search method of *k* nearest neighbours (*kNN).*

6 MODELING THE ALGORITHMS OF FORMING GENERALIZED NOTIONS IN THE PRESENCE OF NOISE

The above mentioned algorithms C 4.5 and CART have been used to research the effect of a noise on forming generalized rules and on classification accuracy of test examples. To restore unknown values the methods of nearest neighbours (*kNN*) and choice of average (*MORM*) are used. We propose the combined algorithm "**I**nduction of **D**ecision **T**ree with restoring **U**nknown **V**alues" (IDTUV3) including C4.5, CART and RECOVERY algorithms. The RECOVERY algorithm is used at the presence of examples containing a noise of the type "*absent values*". When absent values of attributes are restored, one of algorithms of notion generalization is used.

Below, we present the pseudocode of the IDTUV3 algorithm.

```
  Algorithm IDUTV3
   Given: $K = K^+ \cup K^-$
   Obtain: decision rules: decision
   tree T, binary decision tree $T_b$.
    Beginning
    Obtaining $K = K^+ \cup K^-$
   Select noise model (absent values or
distorted values ), noise level,
one of three noise entering types
(uniform, in root column, irregular)
     If noise model is "absent values "
     then use RECOVERY
   Perform: introduce noise in K
   Select generalization algorithm
   C4.5 or CART
      If Select C4.5
      then obtain decision tree T
      else Select CART and obtain
      decision tree $T_b$
       end if
  Output decision tree T or $T_b$
  end
     End of IDTUV3.
```

To develop the generalization system, the instrumental environment MS Visual Studio 2008, program language C# has been used. The given environment is a shortened version MS Visual Studio. DBMS MS Access was used to store data sets.

The program IDTUV3 performs the following main functions:

- loads the original data from DB;
- enters different variants of noise in learning and test sets;
- builds the classification model (a decision tree, or binary decision tree) on the basis of the learning sample;
- forms production rules in accordance with the constructed tree;
- recognizes (classifies) objects using a classification model;
- statistics on classification quality is formed.

We present experiment results fulfilled on the following three data groups from the known collection of the test data sets of California University of Informatics and Computer Engineering "UCI Machine Learning Repository" (Merz and Murphy, 1998):

1. Data of Monk's problems;
2. Repository of data of the StatLog project:
   - Australian credit (Austr.credit);
3. Other data sets (from the field of biology and juridical-investigation practice).

Below the results of experiments for the problem of Monk-1 are produced. The learning sample was made up of 124 examples (62 positive and 62 negative examples), the size of which composes 30% from the whole example space (432 objects).

The set of test examples includes all examples (216 positive and 216 negative examples). Each object is characterized by 6 informative attributes. The class attribute has two different values.

In tables 1, 2 classification accuracy, obtained under the given noise level, is represented for the given generalization algorithm and the particular way of entering noise.

Table 1: Classification results for examples with noise ("distorted values") by entering a noise to a learning sample.

Algorithm and way of entering a noise	Classification accuracy of "noisy" examples, %			
	No noise	Noise 10%	Noise 20%	Noise 30%
C4.5, uniform	85,13	83,98	75,15	79,48
C4.5, irregular	85,13	83,24	77,33	73,31
CART, uniform	81,24	78,34	74,19	72,24
CART, irregular	81,24	77,05	70,73	70,63

Table 2: Classification results for examples with noise ("distorted values") by entering a noise to a test sample.

Algorithm and way of entering a noise	Classification accuracy of "noisy" examples, %			
	No noise	Noise 10%	Noise 20%	Noise 30%
C4.5, uniform	85,13	84,41	76,72	76,96
C4.5, irregular	85,13	81,63	71,97	67,53
CART, uniform	81,24	79,91	75,19	73,68
CART, irregular	81,24	78,59	73,25	72,04

In the case of using the model "absent values" (Tables 3, 4) in each cell there are allocated 2 values: upper – by restoring unknown values on the method of "nearest neighbor" (neighbor number – 3), down - omission of examples with unknown values.

Let us consider the influence problem of a noise entering method on the work of generalization algorithms. Below the experiments are presented in which the model "distorted values" was used and the algorithm C4.5 was chosen. A noise was entered in test samples by each from three ways of entering a noise: uniform, in a root attribute, irregular. Results are given in Table 5 for problems of Monks 1, 2 and 3 and for data collections Glass and Australian Credit.

Table 3: Classification results for examples with noise ("absent values") by entering a noise to a learning sample.

Algorithm and way of entering a noise	Classification accuracy of "noisy" examples, %			
	No noise	Noise 10%	Noise 20%	Noise 30%
C4.5, uniform	85,13	85,35	84,39	79,95
	85,13	82,49	82,71	71,41
C4.5, irregular	85,13	85,12	78,50	77,32
	85,13	79,32	69,91	68,09
CART, uniform	81,24	82,17	76,83	73,14
	81,24	78,28	73,51	70,09
CART, irregular	81,24	82,93	75,68	71,43
	81,24	77,77	68,48	66,54

Table 4: Classification results for examples with noise ("absent values") by entering a noise to a test sample.

Algorithm and way of entering a noise	**Classification accuracy of "noisy" examples, %**			
	No noise	Noise 10%	Noise 20%	Noise 30%
C4.5, uniform	85,13	82,39	74,54	77,72
	85,13	84,38	71,29	73,85
C4.5, irregular	85,13	78,94	71,76	74,42
	85,13	78,06	65,78	66,71
CART, uniform	81,24	77,37	70,33	74,76
	81,24	76,51	64,47	64,81
CART, irregular	81,24	79,70	72,45	75,13
	81,24	78,81	66,41	66,24

Let's define in the Table 5 the methods of entering a noise as follows: *u* – uniform, *r* – in root attribute, *i* – irregular.

We can make the following conclusions. A noise in DBs influences essentially on the classification accuracy and on generalization algorithms as a whole.

The noise entered into a test set has essentially larger influence on the classification accuracy than a noise entered in a learning set (on the average up to 5 – 6% at entering a noise up to 30%).

With increasing a noise level, the irregular way of entering a noise has essentially larger influence on the classification accuracy than the uniform way of entering a noise (on the average up to 3 – 4% at entering a noise up to 30%).

Table 5: Classification results for examples with noise ("distorted values") by entering a noise to a test sample.

Data set and method of entering a noise		Classification accuracy of "noisy" examples, %				
		No noise	Noise 5%	Noise 10%	Noise 15%	Noise 20%
MONKS 1	*u*	82,3	83,53	82,74	83,06	78,14
	r		81,63	81,12	79,73	76,71
	i		83,49	81,89	82,01	76,98
MONKS 2	*u*	88,54	84,43	82,15	79,35	73,5
	r		83,15	79,36	75,28	65,98
	i		82,71	80,82	74,68	68,12
MONKS 3	*u*	85,44	82,35	83,78	79,2	75,89
	r		82,24	80,46	79,81	70,52
	i		82,13	81,59	81,37	71,77
GLASS	*u*	70,35	68,93	67,03	62,93	59,15
	r		65,34	63,71	61,26	55,74
	i		64,48	64,52	63,68	56,61
AUSTR. CREDIT	*u*	83,31	82,73	80,57	73,19	69,41
	r		79,34	75,61	73,33	62,01
	i		82,14	77,49	74,07	63,71

Under growth of a noise level, "distortion model" sometimes is able to increase the classification accuracy.

The method of "nearest neighbours" gives a better classification accuracy in comparison with exclusion from a sample of examples with unknown values (on the average up to 8% under a noise level up to 30%).

The dependence of classification accuracy on a noise level at different variants of entering a noise is close to linear.

From three ways of entering a noise, the most influence on the classification accuracy has entering a noise in the root node.

7 CONCLUSIONS

In the paper the problem of inductive concept formation was considered and means of its solution in the presence of noise in original data were researched.

The new noise model in DB tables was offered, in consequence of which there is provided irregular entering a noise into informative attributes of learning and test samples. Machine experiments on researching the noise influence on work of generalization algorithms C4.5 and CART are produced. The joint algorithm IDTUV3 allowing to process learning samples containing examples with noisy values, have been suggested.

The obtained results of modeling have shown that the algorithms C4.5 и CART in the combination with the restoring algorithm allow to handle data efficiency in the presence of noise of a different type.

The system of building generalized concepts using the obtained results has been developed and the program complex has been implemented.

The paper has been written with the financial support of RFBR (No. 11-07-00038a, 12-01-00589a).

REFERENCES

Vagin V., Golovina E., Zagoryanskaya A., Fomina M., 2008. Exact and Plausible Inference in Intelligent Systems. 2-nd Edition. (in Russian)

Vagin V., Eremeev A., 2009. Methods and Tools for Modelling Reasoning in Diagnostic Systems. *ISEIS 2009, Proceedings of the 11th International Conference on Enterprise Information Systems,* Milan, Italy.

Quinlan J., 1986. Induction of Decision Trees. In Machine Learning, 1.

Quinlan J., 1996. Improved Use of Continuous Attributes in C 4.5. In *Journal of Artificial Intelligence Research*, Vol.4.

Breiman L., Friedman J., Olshen R., Stone C., 1984. Classification and Regression Trees. Wadsworth, Belmont, California.

Mookerjee V., Mannino M., Gilson R., 1995. Improving the Performance Stability of Inductive Expert Systems under Input Noise. *In Information Systems Research* Vol.6 (4).

Vagin V., Fomina M., 2011. Problem of Knowledge Discovery in Noisy Databases. I*n Int. J. Mach. Learn. & Cyber. Vol.2. Springer Ferlag*, Berlin.

Quinlan J., 1986. The effect of noise on concept learning. In Machine Learning Vol. II , Chapter 6. Palo Alto, CA: Tioga.

Vagin V., Fomina M., 2010. Methods and Algorithms of Information Generalization in Noisy Databases. In *Advances in Soft Computing. 9th Mexican Intern. Conference on AI, MICAI 2010*, Pachuca, Mexico, Proceedings, Part II. / Springer Verlag, Berlin.

Vagin V., Fomina M., Kulikov A., 2008. The Problem of Object Recognition in the Presence of Noise in Original Data. In *10th Scandinavian Conference on Artificial Intelligence* SCAI 2008.

Merz C, Murphy P., 1998. UCI Repository of Machine Learning Datasets. Information and Computer Science University of California, Irvine, CA 92697-3425 http://archive.ics.uci.edu/ml/.