Abstract – The table organized data can be analyzed by various algorithms; some of them are capable of generating IF THEN decision rules which comprises of condition attributes and decision attributes. However, it is possible to reduce the set of condition attributes but without information loss. By analysis of the condition attributes set and cuts histogram obtained by discretization and rule consistency, it is possible to choose condition attributes. This paper gives some directions and the practical example.

I. INTRODUCTION

This paper presents the process of attribute selection in order to generate IF THEN rules from table-organized data. The problem is recognized by the CRISP-DM standard [1], which clearly separates Data Understanding from Data Preparation. Data Understanding is an activity that uncovers the very essence of the data, while the Data Preparation includes more activities dealing with lost information, simpler and better data processing. The problem of choosing data for research is especially present in the field of Big Data.

There are situations when the result of data classification obtained by IF THEN rules set is good. However, if we analyze the consistency of this set, we can conclude that there are situations when the rules are not precise enough. If the OR logical operator is included in the THEN part of the rule, the rule is inconsistent and imprecise, while classification is often good. The possibilities of further detailed selection of condition attributes will be considered based on Rough Set Theory [2], and by attribute histogram analysis as well.

II. ROUGH SET THEORY

The Rough Set Theory has been developed having in mind data analysis and information systems [2]. The basic purpose of these sets is the approximation of unfamiliar knowledge using the familiar one [3]. Based on the principle of indiscernibility relation of objects and the concept of approximation, this theory enables recognition of inter-dependability between the decision attributes and condition attributes [4].

In the rough set theory, an information table [5], consisting of ordered quadruple $S = \langle U, Q, V, f \rangle$, is defined, where:

- $U$ is the finite set of objects - universe;
- $Q = \{q_1, q_2, \ldots, q_m\}$ is a finite set of attributes;
- $V = \bigcup_{q \in Q} V_q$, where $V_q$ is the domain of attribute $q$ (attribute values);
- $f = U \times Q \rightarrow V$ is the total function, such that $f(x, q) \in V_q$ for each $q \in Q, x \in U$ and is called information function.

Each object $x \in U$ is described by the vector:

$$\inf_q(x) = [f(x, q_1), f(x, q_2), \ldots, f(x, q_m)]$$

which defines the values of object $x$ attributes.

If $P$ is a non-empty subset of attribute set $Q$, then the relation $I_P$ is defined on the objects from the universe $U$, in the following way:

$$I_P = \{(x, y) \in U \times U : f(x, q) = f(y, q), \forall q \in P\}$$

Relation (2) is called indiscernibility relation. If $(x, y) \in I_P$, we say that the objects $x$ and $y$ are $P$-indiscernible. The indiscernibility relation is the equivalence relation and it generates the partitions – equivalence classes. [6]. The family of equivalence classes, generated by $I_P$, are marked with $U/I_P$. Equivalence classes generated by relation $I_P$ are called $P$-elementary sets, and the equivalence class containing the object $x \in U$ is marked with $I_P(x)$ or $[x]_P$.

For the set $X$, a non-empty subset of $U$, called a rough set, and for $\emptyset \neq P \subseteq Q$, the following are defined:

$$\underline{P}(X) = \{x \in U : I_P(x) \subseteq X\}$$

$$\overline{P}(X) = \bigcup_{x \in X} I_P(x)$$

$P(X)$ is the $P$-lower approximation, which means that the objects certainly belong to $X$, and $\overline{P}(X)$ is the $P$-
upper approximation, which means that the objects may belong to the set \( X \).

\[ P \text{-boundary of subset } X \text{ in } S \text{ is defined as follows:} \]

\[ Bn_p(X) = \bar{P}(X) - P(X) \]  

(5)

Graphical interpretation of P-boundary is shown in Fig. 1. It cannot be said with certainty that the elements from the boundary region belong to the set \( X \) – so it is said that they are the elements of the set \( X \) approximation [8].

This research will use simple calculation of consistency and reduct sets, which are used for IF THEN rules generation. Consistency is defined on the basis of generalized decision function \( \partial \) in the rough set theory. Moreover, a direct implication of inconsistent table can be observed: it is the decision rules which in IF part have the same conditions, and in THEN part various decisions. Information reduct, or just reduct is intuitively recognized as a minimal subset of condition attributes which keep the discernibility between the objects [9]. The reduct set includes only these condition attributes which are sufficient for evaluation of the decision attribute value.

The variation of the Johnson’s algorithm [10] will be used for calculating the minimal simple implicants of Boolean function, so that IF THEN classification rules will be generated by single reduct set. Johnson’s algorithm is a simple greedy heuristic algorithm, often applied to discernibility functions to find a single reduct. The algorithm first sets the current reduct candidate, to the empty set. Then, it evaluates each conditional attribute appearing in the discernibility function according to the heuristic measure. This measure is usually a count of the number of appearances an attribute makes within clauses. The algorithm adds the attribute with the highest heuristic value, removing all clauses in the discernibility function containing this attribute at the same time [11].

After reduct set calculation, it is possible to determine rules of the IF \( a \) THEN \( b \) form. Here \( a \) denotes a conjunction of attributes from some reduct set and their corresponding values: attribute value or subinterval obtained through discretization. The \( b \) consists of decision attribute and corresponding value.

III. HISTOGRAM

Histogram is a well-known mathematical tool for graphical presentation of object distribution, i.e. frequency of certain data values. Its advantage in presenting data is firstly in the fact that the graphical presentation can show the center of distribution, the spread of distribution, shape of data distribution, as well as noticing of some unusual data characteristics. If a distribution has only one peak, then the middle of such a distribution is usually at the peak. The shape of distribution is usually used in recognizing the data patterns [12]. Unusual characteristics often refer to certain breaks in histogram, as well as in isolated values (Fig. 2). Some distributions can be similar to a mathematical distribution, while a significant data set often comes in the form of an irregular multimodal distribution (Fig. 3).

Figure 1. P-boundary of set

Previous to rough set theory application, the data values need to be prepared so that all values are discrete and the consequent histogram is a suitable for data analyses.

IV. THE RELATION OF HISTOGRAM AND DATA DISCRETIZATION

In the domain of Data Mining, numerous Machine Learning methods can work only with the discrete attribute values. This is the reason why it is necessary to transform the continuous attribute values into the discrete ones, before machine learning process. This process of data discretization is an essential task in preprocessing of data. The result of discretization is the set of cuts by which the data are classified into intervals [13].

A. The relation of data distribution and discretization algorithm

The relation of data distribution and large databases has been observed within the algorithm of maximal discernibility, by using median [5]. Paper [14] suggests the usage of the incremental algorithm for data streams, in order to modify the data distribution. The algorithm for estimating a histogram density based on the MDL principle (MDL Histogram Density Estimation) uses the entropy to generate the histograms for data regularity. The consequent analysis results enable the discretization [15]. Therefore, the MDL based algorithm discovers patterns in the data histogram.
B. Histogram segmentation

Histogram segmentation definition is related to recognition of the thresholds of multimodal distribution. Firstly, the segmentation of multimodal distribution is done by ‘smoothing out’ the existing histogram – so that the distribution function becomes a smooth curve, as shown in Fig. 4 (the figure was taken from Paper [16]). Secondly, based on the cross-section of these smooth curves the threshold point is obtained, dividing the histogram into two parts.

![Original histogram (left) vs smooth histogram (right)](image)

Figure 4. Original histogram (left) vs smooth histogram (right)

Appropriate histogram segmentation example is described in [16]: Fig. 5 shows the blood cells and the histogram with two thresholds (the figure was taken from Paper [16]). In detail, Fig. 5 shows the picture of blood cells, histogram, smooth curve obtained through the histogram and the picture obtained through histogram segmentation. The first part of segmentation of the interval (0, 128) shows the picture of blood cells. The second part of the interval segmentation (128, 184) shows the blood plasma. Finally, the segmentation of the interval (184, 250) shows the cell membranes.

![Blood cells](image)

Figure 5. The picture of blood (a), histogram (b), smooth histogram (c) and the results obtained through histogram segmentation, blood cells (d), blood plasma (e) and cell membranes (f)

Based on the histogram analysis, the unimodal parts of multimodal distribution can be recognized. In the case of no-peak histograms or the ones with only one distinct peak, the segmentation would not be possible.

C. The idea of choosing the condition attributes on the basis of histogram and cuts

In [13] the significance of histogram segmentation in the processes of discernibility-based discretization is described. The cuts obtained through the algorithm of maximal discernibility, which are closest to the histogram segmentation thresholds, are very important for the preservation of discernibility. Based on this, the algorithm of approximate discretization APPROX MD was developed [13]. It uses the cuts obtained through the maximal discernibility algorithm, which are closest to the multimodal histogram segmentation thresholds. The aim is to define the most significant cuts for all condition attributes.

Consideration of process of choosing the condition attributes for generating IF THEN rules is the extension of the process described in [13]. Since the reduct of the condition attribute set distinguishes those attributes which can describe the entire set, the subject of the research is the analysis of the histogram of the attributes which make up the reduct.

This paper will give the basic directives for the analysis of histogram and cuts obtained from reduct set. The directives are:

- the existence of the histogram thresholds,
- the position of cuts on the multimodal distribution histograms,
- the position of cuts on the unimodal distribution histograms,
- the comparison of the reduct of the entire information table with the reduct of the modified information table obtained through the reduction of certain condition attributes set,
- the comparison of the reduct of the entire information table with the reduct of the modified information table obtained through the enlargement of certain condition attributes set.

Mandatory part of the reduct analysis is the observation of consistency of the IF THEN rules set.

V. Example

The histograms of condition attributes will be shown on the example of Iris database [17]. The database contains 3 classes, 150 instances while each class refers to a type of the flower iris. As a result, a discretization has been done by the maximal discernibility algorithm and a reduct was calculated by the variation of the Johnson’s algorithm. The reduct consists of third and fourth attribute, which satisfies multimodal histogram distribution. First two condition attributes do not satisfy typical multimodal distribution. The histograms of all the condition attributes are presented on Fig. 6 – Fig 9. The cuts are shown on the ordinates by wide black vertical lines.

![Data distribution of the 1. attribute sepal length](image)

Figure 6. Data distribution of the 1. attribute sepal length, with the cut obtained by the maximal discernibility algorithm
position of the cuts in relation to the thresholds of segmentation of multimodal distribution. The influence of patterns which can be discovered by investigation of the precise IF THEN rules will be obtained.

length) is omitted, the reduct \( \{a_1, a_2, a_3, a_4\} \) is obtained, the third attribute, the inconsistency of the IF THEN rules while the inconsistency increases. In the case of omitting distribution threshold.

of the third attribute overlaps with the multimodal distribution would also contribute to the research of relations between the attribute value data distribution and distribution on quality of decision rules. The correlation attributes, would answer the question on influence of data values, as was shown by the authors who have researched the picture histogram segmentation. The existence of thresholds describes the natural line between the rule consistency and attribute distribution on quality of decision rules. The correlation between the rule consistency and attribute distribution would be the feedback on certain thresholds’ and cuts’ importance. The comparison of various reducts obtained through a range of attributes on the basis of data distribution would also contribute to the research of relations between the attribute value data distribution and the reducts.

Condition attributes histograms have certain hidden patterns which can be discovered by investigation of the position of the cuts in relation to the thresholds of segmentation of multimodal distribution. The influence of multimodal distribution on the reduct set should be researched on the bigger number of the databases.

ACKNOWLEDGMENTS

This paper has been supported by the Ministry of Education and Science of Republic of Serbia, within the project TR32044 “The Development of Software Tools for Business Process Analysis and Improvement”, 2011-2017.

REFERENCES