Construction of Intelligent Systems Based on Knowledge Discovery in Datasets

Viktor Krasnoproshin
Faculty of Applied Mathematics and Computer Science
Belarussian State University
Minsk, Belarus
Email: krasnoproshin@bsu.by

Vadim Rodchenko and Anna Karkanitsa
Faculty of Mathematics and Informatics
Yanka Kupala State University of Grodno
Grodno, Belarus
Email: rovar@grsu.by, a.karkanica@grsu.by

***Abstract*—The original method of intelligent systems construction based on technology of knowledge discovery in databases is considered. To form the knowledge base of an intelligent system, it is proposed to abandon the classical approach based on the formalization of expert knowledge in favor of an alternative approach aimed at identifying interpretable empirical patterns using Data Mining methods.**

***Keywords*—Artificial intelligence, knowledge base, data mining, precedent-based learning**

I. Introduction

The contemporary development of information systems is closely related to artificial intelligence technologies [1], [2]. Therefore, the development of methods and technologies aimed at construction of Intelligent Systems (IS) and their components is one of the key challenges of computer science [3], [4].

An intelligent system is a technical (or software) system capable of solving creative problems in a particular subject domain (SD). At the same time, knowledge about the SD is stored in the memory of the system itself. The IS structure traditionally consists of three major subsystems - a knowledge base (KB), a decision-making mechanism, and an intelligent interface [5], [6].

The knowledge base is the central component of an intelligent system. It contains structured information about the subject domain as a set of facts and rules. The study of models and methods related to the extraction, structuring and representation of knowledge is dealt with by one of the sections of computer science called knowledge engineering [7].

Deductive and inductive learning methods are used to build knowledge base. Deductive methods provide for the formalization of expert knowledge for the purpose of their further placement in the knowledge base. An example of intelligent systems built on the principles of deduction are expert systems (ES) [8], [9]. Inductive methods are based, as a rule, on the principles of learning from precedents. These methods are aimed at identifying empirical patterns in data and are currently associated with machine learning and data mining [10].

The result of machine learning is an algorithm that approximates the unknown target dependency both on the objects of the training set and on the entire initial set [11], [12]. Data mining methods provide for the detection of previously unknown (practically useful and accessible for interpretation) patterns for decision-making in various fields of human activity [13], [14].

Expert systems (since their appearance in the mid-60s of the twentieth century) have demonstrated their effectiveness in such areas of human activity as medicine, chemistry, transport logistics, nuclear engineering, etc. However, today the concept of expert systems is in a serious crisis. The traditional approach to ES construction, unfortunately, does not fit well with the data models of the classical database theory. This makes it impossible to effectively use modern industrial DBMS to form knowledge bases of ES [15].

Currently, the dominant position in the development and construction of intelligent systems is occupied by machine learning and artificial neural networks. At the same time, the task of high-quality learning from precedents is singled out as a central problem [16]. Traditionally, this problem is reduced to solving an optimization problem: it is required to build an algorithm that would best approximate the unknown target dependency, both on the elements of the training set and on the entire set.

The paper proposes a new approach to building intelligent systems based on the technology of knowledge discovery in databases and the original implementation of the data mining stage.

II. Strategies for Intelligent Systems Construction Based on Deductive and Inductive Learning Methods

An intelligent system is a technical (or software) system capable of solving creative problems in a specific subject domain based on knowledge. Knowledge is a set

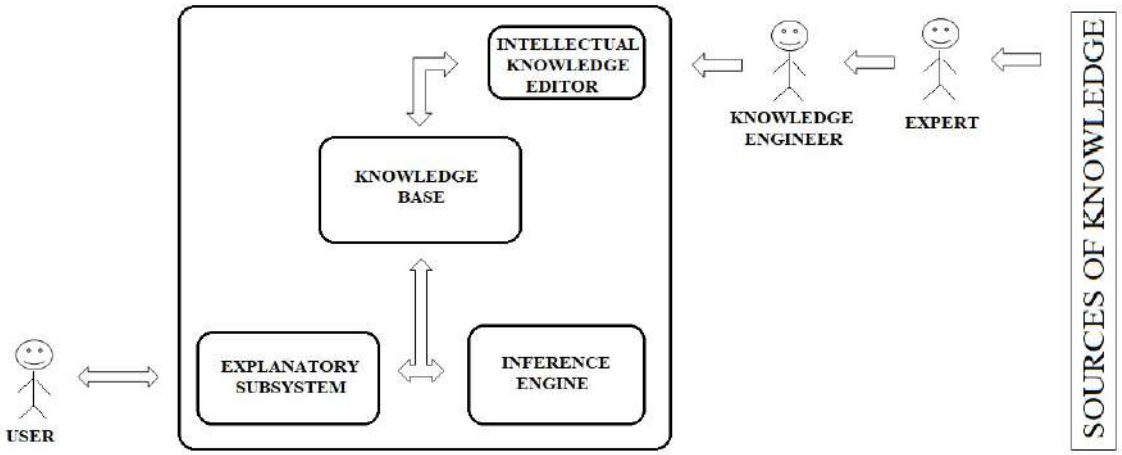

Figure 1. Expert System.

of facts, patterns and heuristic rules necessary to solve a given problem.

To build intelligent systems, two major strategies based on the principles of deduction or induction are used (Fig. 2).

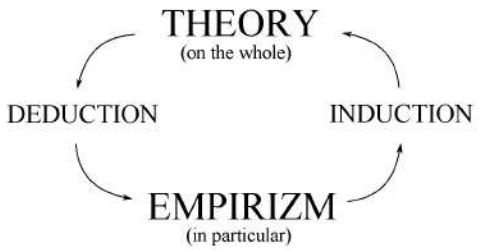

Figure 2. Deduction/Induction process.

Classical representatives of IS built on the basis of deduction are expert systems. The history of their application has more than fifty years. ES is an intelligent system built on knowledge about the subject domain received from experts or from other information sources [17], [18].

Traditionally, the architecture of an expert system consists of three main components: a knowledge base, an inference engine, and an intelligent interface.

The central component of ES is a knowledge base which contains a set of facts and rules necessary to solve the problem. The knowledge base is filled by an expert and a system analyst (knowledge engineer). The expert determines the knowledge base composition, ensures the completeness and correctness of the entered knowledge. The analyst helps the expert to identify and structure knowledge, choose an adequate model for knowledge representation and an effective inference mechanism (Fig. 1).

At present, a separate direction of artificial intelligence has been formed - knowledge engineering. This direction is directly related to the theoretical and practical problems of designing and developing knowledge bases. Knowledge engineering also covers the issues of extracting (or acquiring) knowledge, its structuring and formalization [19].

The process of extracting knowledge is a time and load-consuming procedure. As a result of its execution, the analyst should build an adequate model of the subject domain that experts use for decision-making.

The inference mechanism implements a generalized procedure for searching the solution of the problem. In accordance with the user's need and on the basis of the knowledge base, a chain of reasoning is built, leading to a specific outcome.

The intelligent interface of the system provides a dialogue of the analyst and the user with the expert system. Using the knowledge subsystem, the analyst creates and edits the knowledge base. The user, through the subsystem of explanations, can ask the expert system questions and receive answers from it, can form and analyze chains of reasoning.

The accumulated experience of expert system developers allows us to state that the process of knowledge extraction remains the "bottleneck" in the construction of applied ES. The need for interaction between an expert and a data analyst excludes the possibility of implementing an automatic mode of knowledge base formation and reduces the objectivity of knowledge.

Currently, most intelligent systems are built on the basis of inductive learning methods (learning from precedents) [20], [21].

At the moment, there is a situation where learning from precedents is implemented on the basis of methods and technologies of machine learning and artificial neural networks (Fig. 3). The knowledge base of the intelligent system is formed as a result of the joint work of data analyst and ML-engineer.

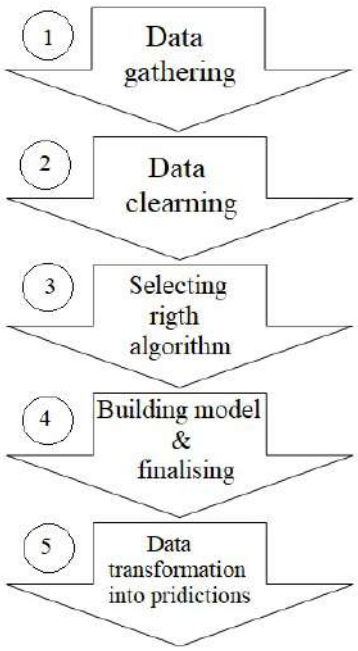

Figure 3. Machine Learning Process.

The work of data analyst begins with the identification and formulation of relevant applied problems in the subject domain. Then data analyst collects information and formulates hypotheses in order to improve certain indicators. And, further, it prepares data for analysis in the form of a training set.

ML-engineer is an expert in the practical application of neural network technologies and machine learning. The task of ML-engineer is to select a model of the problem, select its parameters, select a training method, and, finally, to qualitatively perform the process of model training. Since all this is carried out by the ML-engineer, learning can only be implemented in an automated but not automatic mode. The only useful outcome of learning is the classification algorithm, which is a *«black box»* and it's working mechanism cannot be interpreted. Thus, having spent serious resources on the formation of a training set, as a result, it is only possible to build a classification algorithm, but in no way to expand knowledge about properties of classes and subject domain.

Building intelligent systems (based on inductive methods of learning from precedents) it is proposed to use an alternative approach that allows to opt out of traditional method of learning (as part of solving the classification problem) and which is based on using the idea of the compactness hypothesis.

III. About an Alternative Method of Learning from Precedents

Learning from precedents is based on the identification of empirical patterns in data, and in practice it is performed as part of solving the classification problem [22]. The classification problem has the following statement:

Let X be the set of objects descriptions, Y — the set of numbers (names) of classes. Suppose there is an unknown target dependency $y* : X \to Y$, which values are known only on the objects of the final training set $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. It is necessary to construct an algorithm $a: X \to Y$, that would approximate this target dependency for any object from the initial set X.

At present, the generally accepted scenario for solving the problem can be described as follows:

1. Some model of algorithms $A = \{a : X \to Y\}$ is selected.

2. A loss function $L(y, y')$ is introduced to measure the deviation of the algorithm $(y = a(x))$ for an arbitary $x \in X$ from the correct value $(y' = y^*(x))$.

3. A quality functional $Q(a, X^m)$ is introduced as a value of the average error of the algorithm a on objects of the training set X^m.

4. Within model A, an algorithm that ensures the minimum value of the mean error on the entire sample X^m is constructed.

At least two serious shortcomings of this scenario should be pointed out. Firstly, the choice of an algorithm model, a loss function, and a quality functional is a non-trivial task for the ML-engineer. Secondly, the result of solving the problem is only a classification algorithm, which is actually a *«black box»* (Fig. 4).

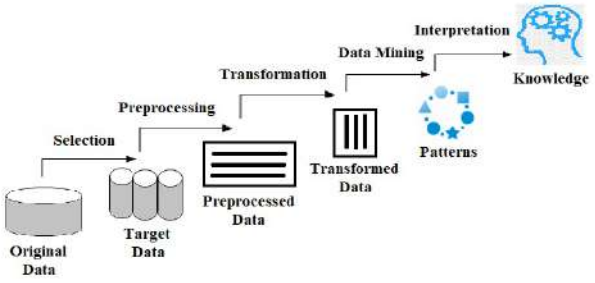

Figure 4. Knowledge Discovery Process.

It is proposed to implement the process of learning from precedents in an alternative way using the idea of the compactness hypothesis.

Learning will be carried out as part of solving the knowledge discovery problem (Fig. 5). The statement of this problem is as follows:

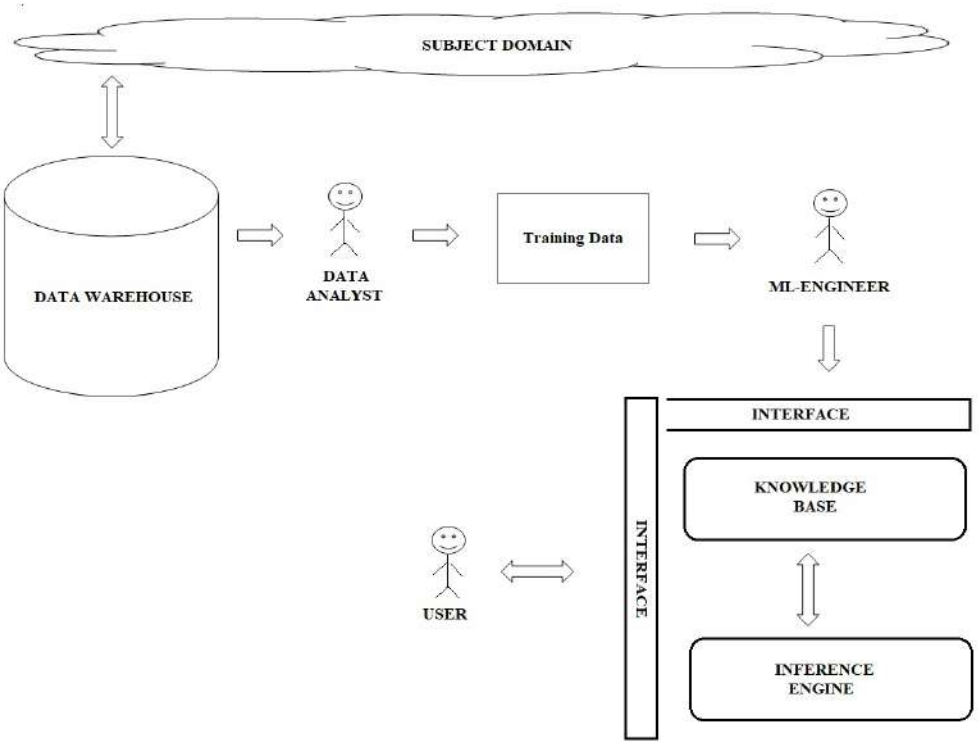

Figure 5. Intelligent System based on Machine Learning.

Let the objects descriptions X, an a priory dictionary of features $F = \{f_1, \ldots, f_n\}$ and classes alphabet $Y = \{y_1, \ldots, y_k\}$ are given. And let the training set $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be formed based on the dictionary F. The set $V = \{v_1, \ldots, v_q\}$, where $q = 2^n - 1$, is the set of possible combinations of features from F. It is required to find such combinations of features from V for which class patterns do not intersect in the corresponding feature subspace.

The algorithm of searching combinations of features from $V = \{v_1, \ldots, v_q\}$ for which class patterns do not intersect in the corresponding feature subspace, is as follows:

Step 1. Select from V a subset $V^+ = \{v_1^+, \ldots, v_n^+\}$, where v_i^+ contains only one feature.

Step 2. For each v_i^+, class patterns are built and their mutual placement is estimated [23].

Step 3. If the patterns do not intersect, then v_i^+ is included in the resulting set $V^* = \{v_1^*, \ldots, v_k^*\}$.

Step 4. The subset $V^+ = \{v_1^+, \ldots, v_n^+\}$ is excluded from the set $V = \{v_1, \ldots, v_q\}$ and set $V^\Delta = \{v_1^\Delta, \ldots, v_p^\Delta\}$ is obtained.

Step 5. All combinations v_i^Δ that contain any combination of $V^* = \{v_1^*, \ldots, v_k^*\}$ are excluded from V^Δ.

Step 6. The next combination v_i^Δ is selected from V^Δ and on its basis a feature space is build.

Step 7. In this feature space, class patterns are build and their mutual placement is estimated.

Step 8. If the class patterns do not intersect, then the combination v_i^Δ is included in the resulting set V^*, and all combinations containing v_i^Δ are excluded from V^Δ.

Step 9. Steps 6-8 are repeated until V^Δ becomes empty.

The result of the algorithm execution will be the set $V^* = \{v_1^*, \ldots, v_t^*\}$, where $0 \leq t \leq q$ and each combination $v_i^* \in V^*$ determines the regularity: *«in the space of combination of features v_i^* classes do not intersect»*.

Since each combination of features from the set V^* defines a decision space where class patterns do not intersect, the construction of classification algorithms does not cause difficulties.

IV. Intelligent Systems Based on Knowledge Discovery in Datasets and OSTIS Technology

An intelligent system built on the basis of the learning method proposed above differs from an intelligent system built on the basis of machine learning in that for formation of a knowledge base an ML-engineer is not required. In addition, the user is provided with the opportunity to analyze the chain of reasoning through the subsystem of explanations. The architecture of an intelligent system based on knowledge discovery in datasets (KDD) is shown in Fig. 6.

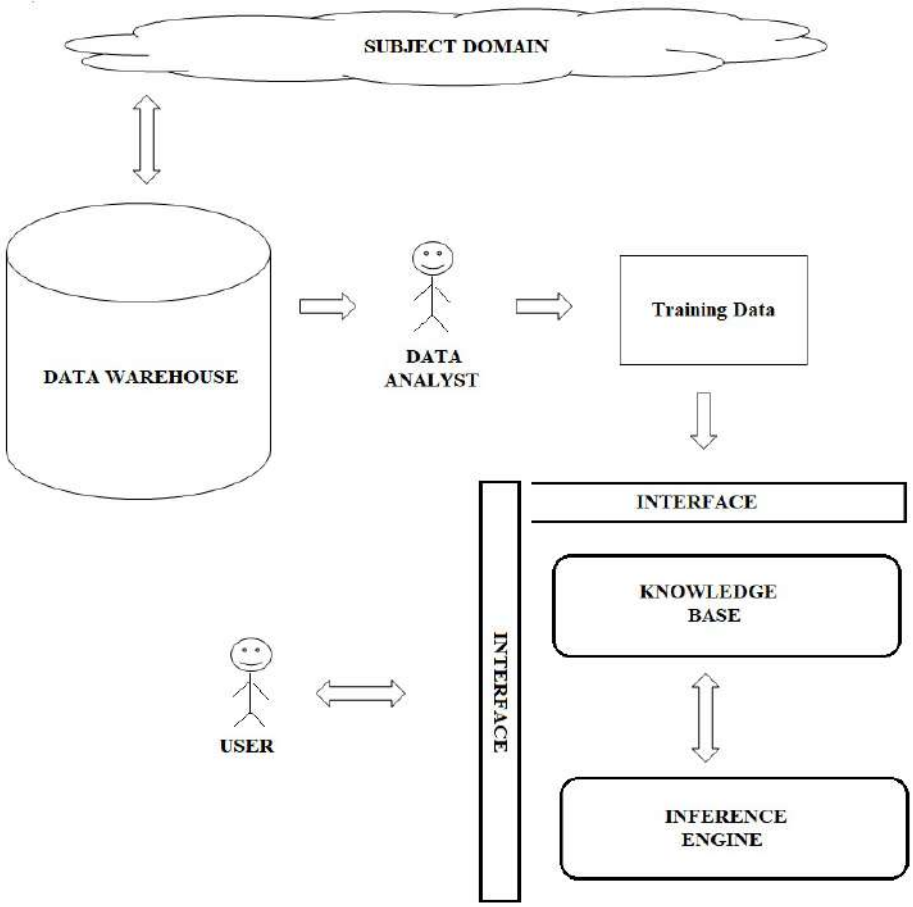

Figure 6. Intelligent System based on Knowledge Discovery.

Data analyst should have a profound knowledge and deeply understanding of processes occurring in the subject domain, be able to correctly set the problem, to collect, process, study and interpret data and, finally, to translate research results into the subject domain language (to make effective decisions). In fact, one of the most important functions of data analyst is the construction of a formal domain ontology.

To design an IS (based on KDD) and represent a formal ontology of the subject domain, it is proposed to use the technology named OSTIS (*Open Semantic Technology for Intelligent Systems*). OSTIS is a comprehensive technology for component-based design of IS [24].

The work of data analyst in forming a training set begins with the construction of an alphabet of classes and an a priori dictionary of features. The usage of OSTIS technology allows to implement a semantic representation of classes, to formalize representation of an a priori dictionary of features and build a training set model using the SC-code (sc-model).

The result of solving the knowledge discovery problem is a set of feature combinations $V^* = \{v_1^*, \ldots, v_t^*\}$. Each combination $v_i^* \in V^*$ defines a pattern of the form *«in the space of combination of features v_i^* classes do not intersect»* and is formally described in the KB using a unified, universal representation language – SC-code.

V. Conclusion

The paper proposes an original method of intelligent system construction based on knowledge discovery. The formation of the IS knowledge base is carried out on the basis of the analysis of training set data. Instead of the generally accepted goal-setting on the construction of a classification algorithm, it is proposed to place an emphasis in the learning process on the study of the classes properties that provide classes distinction.

The original knowledge discovery algorithm is developed which allows to identify automatically combinations of features that provide classes distinction using the features of the a priori dictionary and data of the training set.

It is shown that as a result, it is possible to detect previously unknown, interpreted within the subject domain and practically useful empirical patterns. This knowledge

can be effectively used to build the knowledge base of an intelligent system.

References

[1] Yu. Nechaev, A. Degtyarev Intellektualnyje sistemy: kontseptsii i prilozheniya [Intelligent systems: concepts and applications], Sankt-Peterburg (Saint Petersburg), SPbGU, 2011, P. 269

[2] S. Russel, P. Norving Iskusstvennyj intellect: sovremennyj podkhod [Artificial intelligence: a modern approach], Moskva (Moscow), Izdatel'skii dom Vil'yams, 2006, P. 1408

[3] V. Krasnoproshin, V. Rodchenko, A. Karkanitsa Specialezed KD-agent for knowledge ecosystems. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2021, pp. 59–62.

[4] P. Flach Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh [Machine learning. Science and art of constructing algorithms that extract knowledge from data], Moskva (Moscow), DMK Press, 2015, P. 400.

[5] G. Luger Iskusstvennyj intellect: strategii i metody resheniya slozhnykh problem [Artificial intelligence: structures and strategies for complex problem solving], Moskva (Moscow), Izdatel'skii dom Vil'yams, 2005, P. 864

[6] A. Osrtoukh Intellektualnye sistemy: monografiya [Intelligent systems: monograph], Krasnoyarsk (Krasnoyarsk), Nauchno-innovatsionnyi tsentr, 2020, P. 316

[7] T. Gavrilova, D. Kudryavtsev, D. Muromtsev Ingeneriya znanij. Modeli i metody [Knowledge engineering. Models and methods.], Sankt-Peterburg (Saint Petersburg), Izdatel'skii dom Lan, 2016, P. 324

[8] J. Giarratano, G. Riley Ekspertnye sistemy: printsipy razrabotki i programmirovanie [Expert systems: principles and programming], Moskva (Moscow), Izdatel'skii dom Vil'yams, 2007, P. 1152

[9] C. Naylor Kak postroit' svoju ekspertnuju sistemy [Build your own expert system], Moskva (Moscow), Energoatomizdat, 1991, P. 286

[10] Deductive reasoning. Available at: https://en.wikipedia.org/wiki/Deductive_reasoning// (accessed 2023, Feb)

[11] V. Krasnoproshin, V. Rodchenko Obuchenie po pretsedentam na osnove analiza svojstv priznakov [Learning by precedents based on the analysis of the features], *Doklady BGUIR*, 2017, No. 6, pp. 35–41.

[12] V. Rodchenko Pattern Recognition: Supervised Learning on the Bases of Cluster Structures, Pattern Recognition and Information Processing. PRIP 2016. Communications in Computer and Information Science, Springer, 2017, vol. 673, pp. 106-113.

[13] Data mining. Available at: https://en.wikipedia.org/wiki/Data_mining (accessed 2023, March)

[14] V. Krasnoproshin, V. Rodchenko, A. Karkanitsa Automatic construction of classifiers by knowledge ecosystem agents. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, 2022, pp. 195–198.

[15] V. Ruchkin, V. Fulin Universalnuj iskusstvennuj intellect i ekspertnue sistemu [Universal artificial intelligence and expert systems], Sankt-Peterburg (Saint Petersburg), BHV-Peterburg, 2009, P. 240.

[16] V. Krasnoproshin, V. Rodchenko Klassifikatsiya na osnove prostranstv reshenij [Classification based on decision spaces], Doklady BGUIR, 2019, No. 6, pp. 20–25.

[17] P. Jackson Vvedenie v ekspertnyje sistemu [Introduction to expert systems], Moskva (Moscow), Izdatel'skii dom Vil'yams, 2001, P. 624.

[18] L. Rutkovskij Metody i tekhnologii iskusstvennogo intellekta [Methods and technologies of artificial intelligence], Moskva (Moscow), Telekom, 2010, P. 520.

[19] T. Gavrilova, V. Khoroshevskij Bazy znanij intellektualnykh sistem [Knowledge bases of intelligent systems], Sankt-Peterburg (Saint Petersburg), Izdatel'skii dom Piter, 2000, P. 384.

[20] S. Raschka Python i mashinnoe obuchenie [Python machine learning]", Moskva (Moscow), DMK Press, 2017, P. 418.

[21] Artificial neural network. Available at: https://en.wikipedia.org/wiki/Artificial_neural_network// (accessed 2023, March)

[22] Machine learning. Available at: https://en.wikipedia.org/wiki/Machine_learning// (accessed 2023, March)

[23] V. Krasnoproshin, V. Rodchenko Klasternye struktury i ikh primenenie v intellektualnom analize dannykh [Cluster structures and their applications in data mining], Informatika [Informatics], 2016, No. 2, pp. 71–77.

[24] V. Golenkov, N. Guliakina, I. Davydenko, D. Shunkevich, A. Eremeev Ontologicheskoe proektirovanie gibridnykh semanticheskikh system na osnove smyslovogo predstavleniya znanij [Ontological design of hybrid semantically compatible intelligent systems based on sense representation of knowledge]. *Ontologiya proektirovaniya [Ontology of designing]*, 2019, No. 1, pp. 132–151.

Построение интеллектуальных систем на основе Knowledge Discovery in Datasets

Краснопрошин В. В., Родченко В. Г., Карканица А. В.

Рассмотрен оригинальный способ построения интеллектуальных систем на основе технологии knowledge discovery in databases. Для формирования базы знаний интеллектуальной системы предложено отказаться от классического подхода, основанного на формализации знаний экспертов, в пользу альтернативного, направленного на выявление методами Data Mining интерпретируемых эмпирических закономерностей.

Received 20.03.2023