

УДК 004.822:004.89:004.855

doi: 10.15622/rcai.2025.004

ОРГАНИЗАЦИЯ СОДЕРЖАТЕЛЬНОГО ДОСТУПА К СИСТЕМАТИЗИРОВАННЫМ ЗНАНИЯМ И РЕСУРСАМ ПО МАШИННОМУ ОБУЧЕНИЮ НА ОСНОВЕ ОНТОЛОГИИ

Ю.А. Загорулько (*zagor@iis.nsk.su*)^{A,B}

Г.Б. Загорулько (*zagor@iis.nsk.su*)^{A,B}

Е.А. Сидорова (*lsidorova@iis.nsk.su*)^{A,B}

И.О. Плотникова (*i.plotnikova1@g.nsu.ru*)^B

^A Институт систем информатики им. А.П. Ершова СО РАН,
Новосибирск

^B Новосибирский государственный университет, Новосибирск

Несмотря на то, что область машинного обучения (МО) активно развивается, она все еще слабо формализована, а наработанные в ее рамках инструменты и ресурсы недостаточно систематизированы. Это не только удлинит период вхождения в область МО, но и затрудняет пользователям эффективный выбор необходимых для решения их задач инструментов и ресурсов. Такое положение дел в области МО вызывает необходимость разработки информационно-аналитического интернет-ресурса, который обеспечивал бы систематизацию знаний и информационных ресурсов по МО и содержательный доступ к накопленным в этой области инструментам, моделям, методам и наборам данных. В докладе описывается подход к построению такого ресурса, базирующегося на разработанной авторами онтологии машинного обучения.

Ключевые слова: машинное обучение, онтология, информационно-аналитический интернет-ресурс, паттерны онтологического проектирования.

Введение

В настоящее время все больше людей оказывается вовлечено в область машинного обучения (МО) [Mohri et al., 2018]. Среди них – преподаватели и студенты, осваивающие МО, ученые, использующие МО в своих исследованиях, представители индустрии и органов власти, применяющие ме-

тоды МО для решения своих практических задач. Несмотря на то, что область МО бурно развивается, она до сих пор слабо формализована, а разработанные в ее рамках методы, средства и ресурсы недостаточно систематизированы. Это не только удлинняет период вхождения в область МО, но и затрудняет пользователям эффективный поиск и выбор необходимых для решения их задач методов, моделей, инструментов и наборов данных.

Такое положение дел в области МО диктует потребность в интернет-ресурсе, который обеспечивал бы систематизацию накопленных в этой области научных знаний и данных и поддерживал содержательный доступ к ним. На данный момент в сети Интернет присутствует большое число ресурсов, относящихся к области МО.

Самым известным в России ресурсом такого рода является MachineLearning.ru – русскоязычный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных (ИАД). Ресурс строится по принципам Википедии. Сейчас ресурс содержит около 1100 статей на русском языке, предназначенных как для профессиональных аналитиков, так и для студентов и преподавателей. На нем представлены учебные курсы по МО и ИАД, информация о наиболее важных публикациях, конференциях и семинарах, достижениях научных школ России и стран СНГ в области МО и ИАД.

Этот ресурс может быть полезным для студентов в качестве источника учебных материалов по МО и ИАД, а для специалистов в этих областях – в качестве справочного материала и источника информации о достижениях в этой области. Но доступ к представленным на нем знаниям и данным затруднен из-за отсутствия их четкой систематизации. Кроме того ценность ресурса снижается из-за отсутствия информации о наиболее востребованных алгоритмах, фреймворках и доступа к существующим в свободном доступе наборам данных и обученным моделям.

Из зарубежных ресурсов по МО необходимо отметить, прежде всего, веб-платформы компаний Kaggle и Hugging Face.

На веб-платформе Kaggle (<https://www.kaggle.com/>) пользователи могут публиковать наборы данных, исследовать и создавать модели, взаимодействовать с другими специалистами по анализу данных и машинному обучению, а также организовывать соревнования по анализу данных и участвовать в них. На платформе размещены наборы открытых данных, предоставляются облачные инструменты для обработки данных и машинного обучения. На этом же ресурсе размещены и реализованы обучающие ресурсы по МО.

Веб-платформа компании Hugging Face (<https://huggingface.co/>) предоставляет доступ к инструментам и датасетам для создания приложений с использованием машинного обучения. На ней доступны библиотека Transformers, созданная для приложений обработки естественного языка,

и платформа Hugging Face Hub, которая позволяет пользователям обмениваться моделями машинного обучения и наборами данных, а также демонстрировать свою работу. На этой платформе размещено большое количество предварительно обученных моделей, которые поддерживают решение задач в различных модальностях, таких как обработка естественного языка, компьютерное зрение и аудио.

Описанные выше веб-платформы, как и подобные им зарубежные веб-ресурсы, предоставляют ценные данные и инструменты, полезные как для начинающих пользователей, так и для продвинутых разработчиков МО-приложений. Однако они не предоставляют удобного доступа к систематизированным знаниям об области МО и используемым в ней моделям, методам, наборам данных и метрикам.

Для создания интернет-ресурсов, предоставляющих такой доступ, предлагается использовать технологию построения интеллектуальных информационных интернет-ресурсов [Загоруйко и др., 2016], базирующуюся на онтологии предметной области. В связи с этим первым шагом на пути создания такого ресурса является разработка онтологии машинного обучения.

1. Основные понятия и термины области машинного обучения

Прежде, чем приступить к описанию разработки требуемой онтологии МО, рассмотрим базовые понятия и термины области знаний МО, которые обязательно должны быть в ней представлены.

Сначала уточним, что под термином «машинное обучение» мы будем понимать область исследования искусственного интеллекта, связанную с разработкой и изучением статистических алгоритмов, которые могут обучаться на данных и обобщать их на данные, которые они раньше не «видели», и, таким образом, выполнять задачи без явных инструкций [Machine Learning, 2025].

Алгоритмы МО обучают модель на основе примеров данных, известных как данные обучения, для того, чтобы делать предсказания или находить решения, не будучи явно запрограммированными на это.

Модель машинного обучения – это тип математической модели, которая после «обучения» на заданном наборе данных может использоваться для прогнозирования или классификации новых данных. В более широком смысле термин «модель» может относиться к нескольким уровням специфичности, от общего класса моделей и связанных с ними алгоритмов обучения до полностью обученной модели со всеми ее настроенными внутренними параметрами.

Данные в МО представляются в виде датасетов, т.е. наборов данных, используемых для обучения моделей машинного обучения [Mohri et al., 2018]. **Датасеты** состоят из примеров (объектов), представленных набором признаков, а также соответствующих им целевых параметров.

Для облегчения поиска необходимых методов и алгоритмов МО они должны быть систематизированы по различным аспектам, например, по типу МО, по типам решаемых задач и т.п.

Наиболее важным аспектом МО является **тип машинного обучения**. В научной литературе выделяется четыре основных типа МО: обучение с учителем, обучение без учителя, обучение с частичным привлечением учителя и обучение с подкреплением [Mohri et al., 2018].

Важным является также деление методов МО по лежащим в их основе математическим моделям и алгоритмам. По этому аспекту методы МО делятся на **классические методы**, в основе которых лежат статистические модели и алгоритмы, и **методы глубокого обучения**, базирующиеся на использовании нейронных сетей.

Еще одним типом МО, который необходимо выделить, является ансамблевое обучение, представляющее собой технику машинного обучения, основанную на совместном использовании нескольких обученных алгоритмов с целью получения лучшей предсказательной эффективности, чем можно было бы получить от каждого алгоритма по отдельности [Rokach, 2010].

Типовыми задачами машинного обучения являются классификация, регрессия, кластеризация, уменьшение размерности и обнаружение аномалий [Mohri et al., 2018]. Для решения каждого типа задач применяются определенные методы МО, которые в свою очередь применяются для решения различных **прикладных задач** – от диагностики заболеваний до анализа и генерации текстов, изображений, аудио, видео и пр.

Важную роль в оценке алгоритмов машинного обучения играют метрики качества, так как они позволяют определить, насколько хорошо модель работает на данных и какие улучшения ей требуются. В связи с этим в онтологии должны быть представлены все используемые на данный момент метрики, в частности, precision (точность), recall (полнота), F-мера (комбинированная мера) и другие.

Для решения своих задач пользователю необходимо получить доступ не только к методам МО и их реализациям, но и к наборам данных и моделям МО, ранее использованным для решения подобных задач. Следовательно, методы МО должны быть связаны с наборами данных и ранее обученными моделями. В свою очередь модели должны быть связаны с наборами данных, на которых они обучались.

Для моделей и методов глубокого обучения важным аспектом является используемая при их реализации архитектура, которая может включать одну или несколько нейронных сетей. В связи с этим в онтологии должны быть представлены такие архитектуры.

Важно представить в онтологии и информацию о публикациях по МО и информационных ресурсах, которые могут содержать ссылки на описание и реализацию методов и датасетов.

Модели и методы МО используются в каких-то приложениях и при этом работают в каком-то окружении (библиотеки, фреймворки, операционные системы и вычислительные устройства). Эту информацию нужно отразить в онтологии, также как и информацию о персонах и организациях, вовлеченных в область машинного обучения, и различных видах деятельности, выполняемых в области МО.

Таким образом, в онтологии должны быть представлены как понятия, специфичные для области МО, такие как метод (алгоритм), модель, задача, набор данных, метрика, нейросетевая архитектура, окружение, приложение, так и понятия, служащие для описания деятельности, выполняемой в любой научной области: персона, организация, деятельность, публикация, информационный ресурс и др.

2. Разработка онтологии машинного обучения

Для того, чтобы онтология была практически полезной, она должна не только представлять все базовые понятия МО и связи между ними, но и содержать описания соответствующих этим понятиям сущностей, т.е. описания конкретных методов, моделей, наборов данных и т.п.

1.1. Обзор онтологий машинного обучения

На данный момент разработано несколько онтологий, так или иначе относящихся к области машинного обучения.

Прежде всего, стоит отметить разработанную консорциумом W3C онтологическую схему ML-Schema [ML Schema, 2016]. Она предоставляет набор классов, свойств и ограничений, которые можно использовать для представления и обмена информацией об алгоритмах интеллектуального анализа данных и машинного обучения, наборах данных и выполняемых с их использованием экспериментов. Однако ML-Schema не удовлетворяет выдвинутым нами выше требованиям к онтологии МО, та как она не дает полного и целостного представления об области МО. В ней нет описания конкретных методов, алгоритмов, датасетов и задач.

Другая онтология из области МО – SML [Kallab et al., 2023]. Сами авторы позиционируют SML как основанную на онтологии модель для описания семантического машинного обучения. SML описывает модели машинного обучения с помощью понятного человеку и машине словаря, чтобы облегчить понимание, оценку и выбор удобной модели МО для использования в данном контексте. Она позволяет представлять и хранить характеристики и рабочие спецификации уже реализованных моделей

машинного обучения (например, используемые ими алгоритмы и обучающие и тестовые наборы данных, результаты их оценки и т.д.), что должно облегчить и улучшить для пользователя, обладающего ограниченными знаниями в области машинного обучения, выбор модели машинного обучения, наиболее подходящей для данного контекста и решаемой задачи. Таким образом, эта, безусловно, полезная онтология в основном ориентирована на подробное описание уже реализованных моделей МО, но не содержит сведений о многих базовых понятиях МО и описаний соответствующих им конкретных сущностей.

В близкой к МО области – интеллектуальный анализ данных (ИАД) или data mining – также разработан ряд онтологий. Наиболее известная из них – OntoDM [Džeroski et al., 2008] включает определения основных сущностей ИАД, таких как типы данных и наборы данных, задачи ИАД, алгоритмы ИАД и их компоненты (например, функция расстояния), ограничения и т.д.

Другая онтология – Expos e [Vanschoren et al., 2010] позволяет подробно описывать эксперименты по анализу данных, включая контекст эксперимента, метрики оценки, методы оценки производительности, наборы данных и алгоритмы.

Обе приведенные выше онтологии, хотя и содержат понятия, являющиеся общими для обеих предметных областей (МО и ИАД), но в них не представлены многие базовые понятия МО и описания соответствующих им сущностей.

Авторам известна только одна онтология MLOnto [Braga et al., 2020], которая содержательно описывает область знаний МО, т.е. в этой онтологии представлены не только понятия, но и конкретные сущности из этой области, например, конкретные алгоритмы МО (Linear Regression, Support Vector Machine и т.п.) и фреймворки (Keras, PyTorch и др.), используемые при решении задач методами МО. К сожалению, эта онтология включает неполный набор базовых понятий, необходимых для описания области МО. В частности, в ней не представлены такие важные понятия, как модели и задачи МО, наборы данных и метрики оценки качества работы алгоритмов МО. Кроме того, в этой онтологии конкретные сущности представлены не объектами (индивидами), а классами, что противоречит принципам онтологического моделирования и делает невозможным детальное описание конкретных сущностей.

Приведенный небольшой обзор показывает, что на данный момент нет онтологии, которая могла бы одновременно и систематизировать фундаментальные знания области МО, и содержать подробные описания разработанных в ней методов, моделей, инструментов и наборов данных. В связи с этим было принято решение разработать новую онтологию МО.

1.2. Реализация онтологии машинного обучения

Реализация новой онтологии МО была выполнена в соответствии с методологией, описанной в [Загоруйко и др., 2020]. Данная методология предлагает в качестве основы для построения онтологии целевой научной предметной области (НПО) использовать базовую онтологию научных предметных областей, включающую понятия, характерные для большинства научных предметных областей, и понятия, служащие для описания научно-исследовательской деятельности, а также систему паттернов онтологического проектирования (паттернов ОП), предназначенных для описания решений типовых проблем онтологического инжиниринга.

Построение онтологии конкретной НПО сводится к специализации паттернов ОП на эту область, при необходимости – разработке новых, специфичных для рассматриваемой области паттернов, и дальнейшем построении на их основе фрагментов целевой онтологии путем конкретизации базовых, специализированных и специфичных для этой области паттернов.

Рассмотрим пример специализации представленного в базовой онтологии паттерна Метод исследования на область МО.

В верхней части рис. 1 показан паттерн *Метод исследования*, а в нижней части – паттерн *Метод МО*, полученный в результате его специализация. Данные паттерны реализованы как классы на языке OWL. При этом класс *Метод МО* является подклассом *Метода исследования* и наследует все его свойства. В паттерн этого понятия были добавлены новые связи, отражающие его специфику: *использует Набор данных*, *использует Модель МО*. К *Методу МО* применяется *Метод оценки качества*, который *использует* разные *Метрики*. Для удобства систематизации *методов МО* среди них выделяются *Классические методы* и *Методы глубокого обучения*. В отдельный класс (*Ансамблевый метод*) выделены методы, использующие ансамбли методов МО.

Классические методы дополнительно группируются по *типу МО* (обучение с учителем, обучение без учителя, обучение с частичным привлечением учителя и обучение с подкреплением).

Разрабатываемая онтология содержит и новые понятия, характерные именно для этой области. Для таких понятий, как *Модель МО*, *Набор данных*, *Метрика* и др., были разработаны паттерны ОП «с нуля».

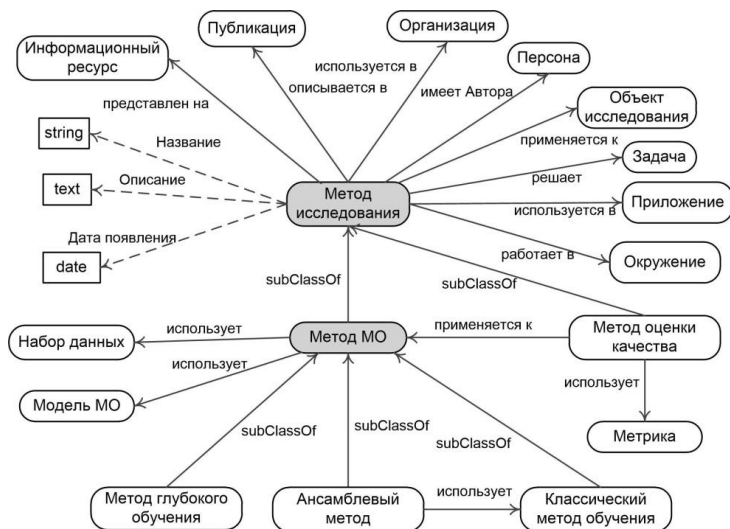


Рис. 1. Специализация паттерна *Метод исследования*



Рис. 2. Паттерн *Модель МО*

На рис. 2 представлен паттерн для описания понятия *Модель МО*. В этом паттерне модель МО связывается с архитектурой, которая используется для реализации модели, с наборами данных, на которых она была обучена, с метриками, используемыми для оценки ее качества, с задачами, для решения которых она предназначена, и с приложениями, в которых она работает. Здесь же могут быть указаны ее авторы и пользователи, а

также публикации о ней и информационные ресурсы, на которых она представлена. Модель МО может основываться на другой модели, т.е. быть получена из нее путем дообучения на каких-то наборах данных. Для представления такой информации вводится атрибутированное отношение “основана на”, связывающее две модели и имеющее атрибут “Обучающий набор данных”. Для этого отношения был разработан структурный логический паттерн.

В связи с наличием огромного количества предварительно обученных нейросетевых моделей в классе *Модель МО* выделяются два подкласса: *Нейросетевая модель* и *Предобученная модель*.

На данный момент существует множество наборов данных, предназначенных для решения задач в различных областях. В связи с этим класс *Набор данных* имеет подклассы: *Набор данных для NLP*, *Набор данных для компьютерного зрения* и *Набор данных общего назначения*.



Рис. 3. Иерархия классов онтологии МО

Базовая онтология НПО и все представленные выше паттерны ОП реализованы на языке OWL. На рис. 3. представлена иерархия классов онтологии МО, построенной в редакторе Protégé.

Включение в онтологию конкретных сущностей выполняется с помощью операции конкретизации паттернов ОП, которая состоит в подстановке в паттерны ОП конкретных значений свойств и добавлении полученных фрагментов в создаваемую онтологию.

3. Реализация ИАИР МО

На основе описанной выше онтологии и упомянутой выше технологии построения интеллектуальных информационных интернет-ресурсов, был реализован ИАИР по машинному обучению (ИАИР МО), который представляет собой доступную через Интернет информационную систему, интегрирующую систематизированные знания, данные и информационные ресурсы из области знаний «Машинное обучение» и обеспечивающую содержательный эффективный доступ к ним.

МАШИННОЕ ОБУЧЕНИЕ



+

Архитектура

+

Географическое место

+

Деятельность

+

Задача

+

Информационный ресурс

+

Метод исследования

+

Метрика

-

Модель машинного обучения

-

Нейросетевая_модель

-

Предобученная модель

-

Предобученная языковая модель

-

Набор данных

-

Набор данных для NLP

-

Набор данных для компьютерного зрения

-

Набор данных общего назначения

-

Область использования

-

Объект исследования

+

Окружение

+

Организация

+

Персона

+

Предмет исследования

+

Приложение

+

Публикация

+

Раздел науки

+

Результат / продукт

+

Событие

Табличное представление

Графовое представление

Свойства объекта

Название	Toronto Book Corpus
Описание	BookCorpus (иногда также называемый Toronto Book Corpus) — это набор данных, состоящий из текстов около 7000 самостоятельно изданных книг, взятых с независимого сайта распространения электронных книг Smashwords. Это был основной корпус, используемый для обучения первоначальной модели GPT компанией OpenAI, и использовался в качестве обучающих данных для других ранних больших языковых моделей, включая BERT от Google. Набор данных состоит из около 985 миллионов слов, а книги, которые его составляют, охватывают целый ряд жанров, включая любовные романы, научную фантастику и фэнтези.
Дата публикации	2015
Размер	7000 книг, 985 миллионов слов
Формат файлов	Текст
Язык	English

Связи объекта

создан для

Задача

Моделирование естественного языка

Обратные связи объекта

обучена на Наборе данных

Модель машинного обучения

BERT

Рис. 4. Интерфейс информационно-аналитического интернет-ресурса по МО

На рис. 4 показана страница этого ресурса. В левой части страницы представлены понятия онтологии, организованные в иерархию по отношению «общее-частное». При выборе конкретного понятия на странице

56

отображается поддерево его понятий-потомков и список объектов, соответствующих этому понятию. При выборе какого-либо объекта из этого списка отображается описание его свойств (атрибутов и связей с другими объектами) в табличном или графическом виде. При этом объекты, связанные с данным, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.

Например, на рис. 4 в табличном виде показано описание набора данных Toronto Book Corpus, из которого можно узнать, что этот набор данных собран в 2015 году, состоит из текстов около 7000 книг, включает 985 миллионов слов, использовался в качестве обучающих данных для большой языковой модели BERT и т.д.

На основе онтологии организуется не только навигация по контенту ИАИР МО, но и содержательный поиск. При этом пользователю доступны два вида поиска: простой и расширенный.

Входом для простого поиска является строка, которая ищется по значениям атрибутов всех объектов, содержащихся в контенте. Результатом простого поиска является список рассортированных по классам онтологии объектов, значения атрибутов которых содержат искомую строку.

При расширенном поиске пользователь формулирует запросы через специальный графический интерфейс, управляемый онтологией. Он может выбрать понятие, к которому относится искомый объект, и задать ограничения, которым должны удовлетворять его свойства.

Заключение

В работе описан подход к разработке интеллектуального научного интернет-ресурса, который обеспечивает содержательный доступ к систематизированным знаниям и данным области МО, тем самым помогая пользователям выбирать конкретные инструменты, методы, модели и наборы данных, необходимые для решения их практических задач. В основе данного ресурса лежит разработанная авторами онтология машинного обучения, в которой формализованы и систематизированы как общие знания об области МО, так и накопленные в этой области методы, модели, инструменты и наборы данных.

На данный момент полностью реализован верхний уровень онтологии МО и выполнено ее частичное наполнение конкретными сущностями. На основе данной онтологии построен рабочий прототип ресурса. В ближайших планах – доведение прототипа ресурса до рабочей версии путем добавления в его контент информации обо всех наиболее важных и популярных инструментах и ресурсах из области машинного обучения.

Кроме того, для поддержки актуальности онтологии предполагается дополнить ресурс модулем, реализующим интеграцию ИАИР МО с наиболее значимыми и популярными ресурсами по МО, в частности, с веб-

платформами компаний Kaggle и Hugging Face. Настройка данного модуля на внешние ресурсы будет осуществляться с помощью соответствующих паттернов онтологического проектирования.

Список литературы

- [Загорулько и др., 2016] Загорулько Ю.А., Загорулько Г.Б., Боровикова О.И. Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии // Программная инженерия. – 2016. – Т. 7, № 2. – С. 51-60. – doi: 10.17587/prin.7.51-60.
- [Загорулько и др., 2020] Загорулько Ю.А., Боровикова О.И. Использование системы разнородных паттернов онтологического проектирования для разработки онтологий научных предметных областей // Программирование. – 2020. – № 4. – С. 27-35. – doi: 10.1134/S0361768820040064.
- [Braga et al., 2020] Braga J., Dias J.L.R., Regateiro F. A Machine Learning Ontology. Preprint. October 2020. – doi: 10.31226/osf.io/rc954
- [Burkov, 2019] Burkov A. The hundred-page machine learning book. Polen: Andriy Burkov, 2019.
- [Džeroski et al., 2008] Džeroski S., Soldatova L., Panov P. OntoDM: An Ontology of Data Mining // In: Proc. 2008 IEEE International Conference on Data Mining Workshops, Pisa, Italy, 2008. – P. 752-760. – doi: 10.1109/ICDMW.2008.62.
- [Kallab et al., 2023] Kallab L., Mansour T., Chbeir R. SML: Semantic Machine Learning Model Ontology / In: Proc. Zhang, F., Wang, H., Barhamgi, M., Chen, L., Zhou, R. (eds.) // Web Information Systems Engineering – WISE 2023, LNCS, 2023. – Vol. 14306. – P. 896-911. – Springer, Singapore. – doi: 10.1007/978-981-99-7254-8_70.
- [Machine Learning, 2025] Machine Learning. – https://en.wikipedia.org/wiki/Machine_learning#cite_note-1, last accessed 2025/06/10.
- [Mitchell, 1997] Mitchell T.M. Machine learning. – McGraw-Hill, New York, 1997.
- [ML Schema, 2016] ML Schema Core Specification: Release 17 October 2016. – <http://ml-schema.github.io/documentation/ML%20Schema.html>, last accessed 2024/08/25.
- [Mohri et al., 2018] Mohri M., Rostamizadeh A., Talwalkar A. Foundations of machine learning, 2nd edn. – The MIT Press, Cambridge, MA, 2018.
- [Rokach, 2010] Rokach L. Ensemble-based classifiers // Artif Intell Rev. – 2010. – Vol. 33. – P. 1-39. – doi: 10.1007/s10462-009-9124-7.
- [Vanschoren et al., 2010] Vanschoren J., Soldatova L. Exposé: An ontology for data mining experiments // In: Proc. Int. workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010). – 2010. – P. 31-46.