How to Cite
Turgunbaev, R. (2024). Machine learning and its use in the automatic extraction of metadata from academic articles. *International Journal of Engineering & Computer Science*, *7*(1), 1-7. https://doi.org/10.21744/ijecs.v7n1.1782

Machine Learning and its Use in the Automatic Extraction of Metadata from Academic Articles

Rashid Turgunbaev
Tashkent Institute of Information Technology, Uzbekistan
Corresponding author email: atmdsmi@gmail.com

Abstract---*This article provides detailed information on machine learning processes, learning methods, supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and the use of machine learning algorithms in the automatic extraction of metadata. Classification and regressive types of supervised learning, including decision trees, decision rules, Naive Bayes classifiers, Bayesian trust networks, nearest neighbor classifiers, linear discriminant functions, logistic regression, support vector machines, artificial neural networks, clustering and dimensionality reduction methods of unsupervised learning, semi-supervised learning, and reinforcement learning methods are also discussed.*

Keywords---*extraction, machine learning, metadata, reinforcement learning, semi-supervised learning, supervised learning, unsupervised learning.*

Introduction

The field of artificial intelligence is typically concerned with the theory and application of computer systems that are capable of performing tasks that require human intelligence. Machine learning is a field of artificial intelligence that has emerged as part of research to create intelligent machines capable of learning (Siswa, 2020). Learning ability is one of the key characteristics of intellectual behavior. Machine learning is the field of artificial intelligence that deals with the identification of patterns based on data and making predictions about new data using them. According to the definition given by Mitchell (1997), a computer program is considered to learn based on experience with a particular task and a certain performance indicator when its performance measured by a given performance improves based on this experience. The main outcome of the machine study is the degree of generalization, i.e., the degree of correct prediction for new data based on the rules studied based on previously similar data of the model. Creating models that can make accurate predictions based on new data is the main goal of machine learning. The transition from explanation to generalization differs from traditional research based on statistical machine learning. Machine learning studies the relevant laws in data and then uses them to make predictions (Al-Jarrah et al., 2015). Machine learning data analysis and knowledge search in databases; automatic creation of knowledge base for expert systems; learn to plan, play games, create quantitative and qualitative models; classification and retrieval of texts; automatic acquisition of knowledge to manage dynamic processes; used in painting, handwriting and automatic speech recognition and other fields. The basic principle of machine learning is to model the processes that create this collected data. As a result of learning from the data, other knowledge manifestations such as rules, functions, relationships, systems of equations, probability distribution, and decision rules appear. Models explain data and are used to support process decisions (Kononenko, 2007).

There are different approaches to machine learning depending on the learning method. According to this classification, machine learning methods are divided into four groups: supervised, unsupervised, semi-supervised, and reinforcement learning (Vieira et al., 2020).

ISSN 2632-945X
Submitted: 27 January 2024 |Revised: 18 February 2024 | Accepted: 09 March 2024

Supervised learning

In supervised learning, the object that the algorithm must predict will be able to refer to the target variable. A more basic goal is to use an algorithm to study the optimal function that achieves the relationship between the input data and the target variable. In supervised learning, the algorithm will have a preview of what the output values should be. The algorithm is taught based on several examples and is allowed to receive feedback depending on the proximity of the prediction to the target. In this context, the study is a repetitive process of forecasting and subsequent modification until the difference between this generated prediction and the target variable is minimized. The performance of the algorithm is measured by comparing the generated predictions with the target values of the new data. Depending on the stability or continuity of the target variable, the supervised learning task is divided into classifying or regressive types. Classification algorithms are designed to predict grouping for observation sets. In regressive problems, the goal is to predict the outcome continually. Some problems that can be solved using a classification algorithm can also be solved using a regressive algorithm by defining the result as a continuous variable.

Classification

Machine learning methods are widely used for classification. Each object is described knowing many attributes. An object can be assigned from a limited set of classes to a specific class (Mair et al., 2000). Object attributes can be discrete or continuous as observable independent variables. A class is a non-observable discrete dependent variable whose value is determined from the values of the corresponding independent variables. The task of the classifier is to determine to which class the object under study is assigned. To define a class, you need to describe a discrete function that moves from the classifier attributes field to the class field. This function can be given in advance or learned from the data. The data may consist of study examples describing previous problems. Different classifiers represent the transfer functions in different ways. Common descriptors include decision trees, decision rules, Naive Bayes classifiers, Bayesian trust networks, nearest neighbor classifiers, linear discriminant functions, logistic regression, support vector machines, and artificial neural networks (Kononenko, 2007).

Decision trees are a non-parametric supervised learning method used for this classification and regression. The main purpose of this method is to create a model for estimating the value of the target variable by studying the rules of decision-making based on the properties of the data. The decision tree can display the selections and their results in the form of a graph in the form of a tree. While the joints of a tree represent an event or a choice, the branches represent the rules or conditions of decision-making. Each decision tree consists of joints and branches. Each joint represents the attributes in the group to be classified and the value that each type of joint can accept (Sharma & Kumar, 2016).

A Naive Bayesian classifier is a probabilistic machine learning model used for classification tasks, based on the Bayesian theorem. According to Bayes' theorem, given the occurrence of event B, the probability of event A is found. Here B is the proof, A is the hypothesis. Assumptions and functions are calculated independently, that is, the presence of one property does not affect another property. The Naive Bayesian classifier is mainly used to classify text (Berrar, 2019).

The nearest neighbour classifier is a machine learning method designed to identify previously unquestioned survey objects when separating two or more target classes (Liu & Özsu, 2009). In general, as with any classifier, it requires exercise information with given characters. The query object inherits characters from the nearest sample object in the exercise set. The set of decisions of one nearest neighbor in the exercise data k is extended to the nearest set of neighbors. The decision-making rule generalizes the characteristics of these k-decision objects in determining the predicted character for the object of inquiry (Gursoy et al., 2017).

The task of the learning algorithm is to calculate the coefficients for the predefined discriminant functions of the structure. The discriminant function is a hypersurface that divides between two classes in this attribute field. Since a hypersurface can only be defined in a continuous hyperspace, all attributes are required to be continuous. If the number of classes is more than two, a separate hypersurface is required for each pair of classes. Discriminant functions can be linear, quadratic, or polynomial. If the discriminant function is linear, the corresponding hypersurface will be divisible between the two classes. Hypersurface coefficients are determined to minimize the level of classification errors. In many cases, Fisher's linear discriminant function is used. It assumes a normal distribution of study samples in each class and maximizes the Euclidean distance between the average samples of both classes. It is calculated for potentially different class differences and finds an acceptable classification boundary between classes (Shimodaira, 2015).

In the algorithm of support vector machines, each data element is defined in n-dimensional space. In this case, n is the number of properties, and the value of the property is the value of the coordinate point. Classification is done by finding the hyperplane that separates the two classes. Hyperplanes are decision limits that help classify data points. The data points corresponding to different sides of the hyperplane will belong to different classes. The size of the hyperplane depends on the number of properties. Base vectors are the data points closest to the hyperplane that affect the location and direction of the hyperplane. The algorithm of the support vector machines has high speed and good performance when working with a limited amount of data (Suthaharan & Suthaharan, 2016).

Algorithms for artificial neural networks mimic a biological neural network by abstracting neural functions into the appearance of a simple neuronal element and can generalize input data and normalize output data. Neurons are interconnected into arbitrarily complex artificial neural networks. Directly connected multilayered neural networks are often used for classification. Neurons are organized in the form of layered cascades, i.e., the input layer corresponding to the attributes, one or more latent layers, and the output layer corresponding to the classes. The main task of the learning algorithm is to determine the weight of the connections between neurons to minimize the error rate of the classification. To classify a new example, its attribute values are presented to the input neurons of the artificial neural network. The weights of these values are measured based on the relationships between neurons, and the sum of their weights in each neuron in the next layer of neurons is calculated. Normalized outcomes in output neurons determine the classification outcome (Kubat & Kubat, 2021).

Regression

As in the case of classification, the regression has a set of objects that are described by several attributes. Attributes are variables that can be tracked independently. The dependent variable is continuous and its value is defined as a function of the independent variables. The task of a regressive predictor is to determine the value of an unobservable continuous variable that is dependent on the object under consideration. Like classifiers, a regressive predictor performs the function of continuous conversion from attribute space to prediction values. This function can be given in advance or learned from previously solved problems. The task of the learning algorithm is to determine the continuous function by learning from the learning set. This function can then be used to predict values for new, previously unseen examples. Regressive predictors differ in the expression of a regressive function. Common regression predictors include linear regression, regression tree, local weight regression, support vector machines for regression, and multilayer directly connected neural networks for regression (Darlington & Hayes, 2017).

The regression tree represents the algorithm in which the target variable is located, and this algorithm is used to predict its value. Regression trees, unlike classification trees, are used when the target variable is continuous. Regression trees are used in forecasting issues. The regression tree is constructed using a binary recursive division process. This process is an iterative process that divides data into sections (tree joints), and the resulting sections into smaller sections.

In the linear regression function, it is assumed that all attributes are continuously evaluated and normalized accordingly. It also implies the existence of a linear relationship between dependent and independent variables. To minimize the sum of the absolute errors of the regressive predictions evaluated in the study algorithm study examples, it is necessary to determine the coefficients of the linear function. An analytical solution or iterative optimization can be applied. The result of the study is a hyperplane that determines the value of the dependent variable for each instance. For the given example, the error is determined by the distance from its dependent variable axis to the hyperplane.

Locally weighted regression is the closest neighbor algorithm adapted to the regression, and the linear regression function is calculated based on the nearest neighbors of the new example. Through this, the linear regression is adapted to local characteristics. In addition to linear regression, the average values of the nearest neighbors or another simple function can be used. Complex functions are not used due to the small number of neighbors.

It is possible to customize the algorithm of the reference vector machines to solve regression problems. The boundary around the regression hyperplane is defined in such a way that the correctly predicted study issues lie within it. The base vectors define the boundary. Support vector machines are aimed at reducing the limit, as well as minimizing prediction errors in study examples. Like the support vector machines used in classification, the criterion function will need to be optimized for the predicted error of the regression variable, as well as the complexity of the criterion function.

Hybrid algorithms for classification combine several types of approaches using their advantages and thus create probabilistic perfect learning algorithms. Locally weighted regression is an algorithm that combines linear regression and the closest neighbor algorithms. Such results are obtained by applying a linear regression algorithm in regression trees. The rationale here is that for a simple regressor to work well, the regression tree will need to have enough space for the attribute space sections. The regression tree should be smaller to provide sufficient examples for normal regression.

Unsupervised learning

Unlike supervised learning, there is no target value in the unsupervised learning task. The main purpose of unsupervised learning is to determine the basic structure of information. There are types of clustering and dimensional reduction of unsupervised learning. Cluster analysis is an analytical method for creating meaningful subgroups from a large sample. The dimensional reduction method is useful when the number of features is much larger than the number of observations, reducing computing power, removing insignificant data, and reducing the risk of overfilling. Basic component analysis, independent component analysis, or autoencoder methods can be used to reduce scalability.

Clustering is the main type of unsupervised learning is the process of grouping similar examples. The purpose of clustering is to find a specific structure in an undefined data set. A cluster is a set of examples that are similar to each other but not similar to the examples in the other cluster. In clustering, the number of groups is not known in advance. The clustering algorithm must meet the following criteria: scalability, working with different types of attributes, identification of clusters of arbitrary shape, minimum knowledge requirements to determine access parameters, ability to work with missing data and interactions, insensitivity to input data order, ease of understanding and use. Clustering algorithms can be classified into hierarchical or fragmentary, exclusive or intersecting, deterministic or stochastic, and incremental or non-incremental types. The hierarchical clustering algorithm is based on the unity between two or more clusters. At the beginning of the process, each issue is treated as a separate cluster. After several iterations of adding clusters, the final clusters are obtained. In partial clustering, examples are divided into clusters. In exclusive clustering methods, examples are grouped separately. That is, if an instance belongs to one cluster, it cannot be added to another cluster. In intersecting clustering methods, each instance can belong to multiple clusters at different membership levels. Differences are a fundamental concept in the definition of a cluster, and the measurement of the degree of dissimilarity between two examples taken from the space of the same attributes is the basis for most clustering processes.

To adequately manage high-dimensional data, their dimensionality will need to be minimized. The dimensionality reduction method is a meaningful expression of dimensionality reduction data with high dimensions. The reduced view of the dimension should correspond to the internal dimension of the data. The internal dimension of data is the minimum number of parameters required to take into account its observed properties (Fukunaga, 2013). Typically, dimensional reduction is accomplished using linear techniques such as key component analysis, factor analysis, and conventional scaling. Recently, nonlinear techniques have been proposed to reduce dimensionality. Unlike traditional linear techniques, nonlinear techniques allow you to work with complex nonlinear data.

Semi-supervised learning

In semi-supervised learning, the target variables are present in only a certain portion of the data. In many cases, obtaining high-quality defined data is expensive and time-consuming. In semi-supervised learning, the model provides integration of undefined data with its supervised learning. This approach is useful for all participants when it is impossible or expensive to address or measure the target variable. Semi-supervised learning algorithms attempt to increase performance by generalizing data related to controlled and unsupervised learning algorithms. Research in the field of semi-supervised learning has focused mainly on classification (Leng et al., 2013). Semi-controlled classification methods are associated with cases where the identified data are scarce. In such cases, the establishment of a reliable controlled classifier can be difficult and occurs in areas where it is difficult or expensive to obtain the identified data. This in turn increases the need for the user to create efficient machine learning algorithms by combining a small amount of defined data with a large amount of undefined data (Tsai et al., 2009). Semi-supervised learning depends on the class of machine learning technique that creates the classifier by combining defined and unmarked data. There are two main problems of semi-supervised learning, which are that undefined data are considered useful for inference and that unspecified data is used to improve the quality of inference. Effective use of unspecified data requires certain assumptions. The function of unspecified data is to limit the scope of potential inference rules, and selecting the correct predictor requires a small amount of defined data. Common assumptions of semi-supervised learning include cluster estimation, multilayer assumption, and compatibility assumption. The cluster estimate has undefined data clusters, and each cluster corresponds to only one class. Separating clusters from undefined data drastically reduces the area for probabilistic classification rules. In the simplest case, it is sufficient for the classification to correspond to one set of data per cluster (Anselmi et al., 2016). In a multilayer estimate, the information lies in or near the small-sized sublayer in the multidimensional space, and the conditional class is placed around the multilayer. The compatibility estimate is based on the fact that the data has a different set of attributes. The assumption here is that the classification obtained using these attributes is also required to be consistent for the undefined part of the data set (Zhou & Belkin, 2014). Semi-supervised learning algorithms include transductive support vector machines, generative models, and self-positioning algorithms (Mahesh, 2020).

Reinforcement study

The goal of reinforcement learning is to create a system that is capable of learning through interaction with the environment. In this type of study, the behavior of the algorithm is shaped by a sequence of rewards and punishments in achieving the goal set by the seeker (Tiwari, 2022). Unlike supervised learning, which has algorithms that use given patterns to model behaviors, reinforcement learning allows the algorithm to operate freely (O'Doherty et al., 2015). Provides an opportunity to find actions that increase rewards and reduce penalties based on trial and error. Consolidating learning is the learning of what to do to maximize rewards, i.e., linking situations to actions. It does not determine which action the learner will perform but will determine which actions will bring the most rewards by performing them. Ideally, actions can affect not only current rewards but also subsequent situations and subsequent rewards. Trial and error search and delayed reward are some of the key features of reinforcing learning (Sutton & Barto, 2018).

Use of machine learning algorithms in automatic extraction of metadata

Automatic extraction of metadata ensures the popularity and widespread use of digital library collections (Turgunbaev, 2021a; Turgunbaev, 2021b; Turgunbaev, 2021c). Machine learning methods provide reliable and flexible automatic extraction of metadata (Turgunbaev, 2021).

Han et al. (2003), propose a method of automatic extraction of metadata from documents using support vector machines. The method based on the classification of support vector machines for the extraction of metadata from the head of academic articles is more effective than other methods. Using this method, each row of the document header is classified into one or more classes. In the previous iteration, the iterative convergence procedure using the predicted class symbols of adjacent rows is used to improve the classification of the rows. Subsequent metadata extraction is done by searching for the best segment boundaries of each row. Data structure templates and field-based word clustering can increase metadata extraction productivity. In addition, the normalization of the correct properties also dramatically increases the productivity of the classification.

Tkaczyk et al. (2015), CERMINE offers a system for automatic extraction of structured metadata from the scientific literature. This system is a comprehensive open-source system that extracts structured metadata from electronic scientific articles. The system is based on a modular workflow, and its loosely connected architecture allows for the evaluation and customization of individual components, easy improvement and replacement of independent parts of the algorithm, and helps to expand the architecture in the future. Many stages of system implementation are based on controlled and uncontrolled machine learning techniques, which in turn facilitate the system to adapt to new document structures and styles. Evaluation of the extraction process performed using a large amount of data has shown good efficiency for most types of metadata.

Safder et al. (2020), proposed a method of extraction based on the in-depth study of algorithmic metadata from full-text academic documents. The development of search engines allows you to efficiently retrieve large amounts of textual information. However, such traditional search methods show a low level of accuracy of the data obtained in most cases. The AlgorithmSeer search engine, designed for algorithms, sees pseudo-codes and superficial text metadata from scientific publications as a traditional document for applying general search engine techniques to. Methods for automatic detection and extraction of sentences containing algorithmic pseudo-codes and related algorithmic metadata using a set of machine learning techniques have been proposed.

Skluzacek et al. (2018), Skluma: An extended metadata extraction system has been proposed for disordered data. The Scluma system has been proposed to mitigate the effect of high-speed data expansion and automate the organization of data repositories. This system automatically processes the target repository and extracts the metadata. The Skluma system is capable of extracting a variety of metadata, including aggregated values derived from data with a built-in structure, named objects and hidden topics within text data, and content within images. The Skluma system performs a comprehensive approximate source when extracting metadata from files (Day et al., 2007). Uses machine learning methods in determining file type, dynamically prioritizes a set of metadata extractors, and applies them in practice, studying metadata based on the relationships between files. The metadata obtained describes the approximate knowledge about each file and can then be used in search and organization processes.

References

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87-93. https://doi.org/10.1016/j.bdr.2015.04.001

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., & Poggio, T. (2016). Unsupervised learning of invariant representations. *Theoretical Computer Science*, *633*, 112-121. https://doi.org/10.1016/j.tcs.2015.06.048

Berrar, D. (2019). Bayes' Theorem and Naive Bayes Classifier.

Darlington, R. B., & Hayes, A. F. (2017). Regression analysis and linear models. *New York, NY: Guilford*, 603-611.

Day, M. Y., Tsai, R. T. H., Sung, C. L., Hsieh, C. C., Lee, C. W., Wu, S. H., ... & Hsu, W. L. (2007). Reference metadata extraction using a hierarchical knowledge representation framework. *Decision Support Systems*, *43*(1), 152-167. https://doi.org/10.1016/j.dss.2006.08.006

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.

Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2017). Differentially private nearest neighbor classification. *Data Mining and Knowledge Discovery*, *31*, 1544-1575.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings.* (pp. 37-48). IEEE.

Kononenko, I. (2007). Chapter 1-Introduction, Editor (s): Igor Kononenko, Matjaž Kukar. *Machine Learning and Data Mining, Woodhead Publishing*, 1-36.

Kubat, M., & Kubat, M. (2021). Artificial neural networks. *An introduction to machine learning*, 117-143.

Leng, Y., Xu, X., & Qi, G. (2013). Combining active learning and semi-supervised learning to construct SVM classifier. *Knowledge-Based Systems*, *44*, 121-131. https://doi.org/10.1016/j.knosys.2013.01.032

Liu, L., & Özsu, M. T. (Eds.). (2009). *Encyclopedia of database systems* (Vol. 6). New York: Springer.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., & Webster, S. (2000). An investigation of machine learning based prediction systems. *Journal of systems and software*, *53*(1), 23-29. https://doi.org/10.1016/S0164-1212(00)00005-4

Mitchell, T. M. (1997). Does machine learning really work?. *AI magazine*, *18*(3), 11-11.

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94-100. https://doi.org/10.1016/j.cobeha.2014.10.004

Safder, I., Hassan, S. U., Visvizi, A., Noraset, T., Nawaz, R., & Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. *Information processing & management*, *57*(6), 102269. https://doi.org/10.1016/j.ipm.2020.102269

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, *5*(4), 2094-2097..

Shimodaira, H. (2015). Discriminant functions. *Learning and Data Note*, *10*.

Siswa, T. A. Y. (2020). The effectiveness of artificial intelligence on education: learning during the pandemic and in the future. *International Journal of Engineering & Computer Science*, *3*(1), 24-30. https://doi.org/10.31295/ijecs.v3n1.195

Skluzacek, T. J., Kumar, R., Chard, R., Harrison, G., Beckman, P., Chard, K., & Foster, I. T. (2018). Skluma: An extensible metadata extraction pipeline for disorganized data. In *2018 IEEE 14th International Conference on e-Science (e-Science)* (pp. 256-266). IEEE.

Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207-235.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tiwari, A. (2022). Supervised learning: From theory to applications. In *Artificial intelligence and machine learning for EDGE computing* (pp. 23-32). Academic Press. https://doi.org/10.1016/B978-0-12-824054-0.00026-5

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, *18*, 317-335.

Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009). Intrusion detection by machine learning: A review. *expert systems with applications*, *36*(10), 11994-12000. https://doi.org/10.1016/j.eswa.2009.05.029

Turgunbaev, R. (2021). Keysga asoslangan fikrlash va uni akademik metama’lumotlarni avtomatik ekstraksiya qilishda tadbiq qilinishi. *Science and Education*, *2*(9), 129-144.

Turgunbaev, R. (2021a). Metadata in Data Search. In *" ONLINE-CONFERENCES" PLATFORM* (pp. 93-96).

Turgunbaev, R. (2021b). The role and importance of metadata in information retrieval. Science and Education, 2(8), 353-359.

Turgunbaev, R. (2021c). Metadata: characteristics, types and standards. Science and Education, 2(5), 167-175.

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2020). Introduction to machine learning. In *Machine learning* (pp. 1-20). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00001-8

Zhou, X., & Belkin, M. (2014). Semi-supervised learning. In *Academic press library in signal processing* (Vol. 1, pp. 1239-1269). Elsevier. https://doi.org/10.1016/B978-0-12-396502-8.00022-X