АНАЛИЗ АЛГОРИТМОВ КОДИРОВАНИЯ КАТЕГОРИАЛЬНЫХ ДАННЫХ

Ш. А. Турсунов, А. Э. Рашидов

Самаркандский государственный университет
имени Шарофа Рашидова,
703004, Самарканд, Узбекистан

УДК 519
DOI: 10.24412/2073-0667-2025-2-65-80
EDN: ALXCCT

Известно, что эффективность искусственного интеллекта, признанного наиболее полезным инструментом во всех сферах, тесно связана с рядом факторов. Один из ключевых факторов — необходимость преобразования входных данных в формат, понятный алгоритмам ИИ, поскольку они основаны на математических операциях. Однако часто встречаются данные, на которых невозможно напрямую выполнять арифметические действия. Удаление таких данных может негативно повлиять на результат, поэтому требуется их преобразование в числовую форму. Существуют различные методы кодирования категориальных данных, и выбор наилучшего из них представляет собой сложную исследовательскую задачу. В данной работе проводится анализ 12 методов преобразования текстовых данных. Рассматриваются преимущества и недостатки каждого подхода, а также проводится их сравнительный анализ с выводами.

Ключевые слова: искусственный интеллект, кодирование категориальных данных, методы кодирования данных.

Список литературы

1. Nazarov, F. M., O'G'Li, Y. S. S., & O'G'Li, E. B. S. Algorithms To Increase Data Reliability In Video Transcription // 2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT), 2022. 1–6. DOI: https://doi.org/10.1109/AICT55583.2022.10013558.

2. Rashidov, A., Akhatov, A., Aminov, I., Mardonov, D., & Dagur, A. Distribution of data flows in distributed systems using hierarchical clustering // In A. Dagur, D. K. Shukla, N. F. Makhmadiyarovich, A. A. Rustamovich, & J. J. Sindorovich, Artificial Intelligence and Information Technologies. 2024. 1st ed., P. 207–212. CRC Press. DOI: https://doi.org/10.1201/9781032700502-34.

3. Zaynidinov, H., Xuramov, L., & Khodjaeva, D. Intelligent algorithms of digital processing of biomedical images in wavelet methods // In A. Dagur, K. Singh, P. S. Mehra, & D. K. Shukla, Artificial Intelligence, Blockchain, Computing and Security. 2023. Vol. 2. 1st ed., P. 648–653. CRC Press. DOI: https://doi.org/10.1201/9781032684994-105.

4. Nazarov, F. M., Yarmatov, S., & Xamidov, M. Machine Learning Price Prediction on Green Building Prices // 2024 International Russian Smart Industry Conference (SmartIndustryCon), 2024. P. 906–911. DOI: https://doi.org/10.1109/SmartIndustryCon61328.2024.10515790.

© Ш. А. Турсунов, А. Э. Рашидов, 2025

5. Rashidov, A., Akhatov, A., & Nazarov, F. The Same Size Distribution of Data Based on Unsupervised Clustering Algorithms // In Z. Hu, Q. Zhang, & M. He (Eds.), Advances in Artificial Systems for Logistics Engineering III. 2023. Vol. 180, P. 437–447. Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-36115-9_40.

6. Rustamovich, A. A., Mekhriddin, N., Fayzullo, N., & Sabharwal, M. Intelligent system of labor market regulation based on the evolutionary modeling of employment // 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2022. P. 2534–2539. DOI: https://doi.org/10.1109/ICAC3N56670.2022.10074149.

7. Rashidov, A., Akhatov, A., & Mardonov, D. The Distribution Algorithm of Data Flows Based on the BIRCH Clustering in the Internal Distribution Mechanism // 2024 International Russian Smart Industry Conference (SmartIndustryCon), 2024. P. 923–927. DOI: https://doi.org/10.1109/SmartIndustryCon61328.2024.10516193.

8. Rashidov Akbar Ergash o'g'li, S. J. S. o'g'li. Selecting Methods Of Significant Data From Gathered Datasets For Research. 2024. DOI: https://doi.org/10.5281/ZENODO.10781255.

9. Rashidov, A., & Madaminjonov, A. Sun'iy intellekt modelini qurishda ma'lumotlarni tozalash bosqichi tahlili: Sun'iy intellekt modelini qurishda ma'lumotlarni tozalash bosqichi tahlili // Modern problems and prospects of applied mathematics, 2024. N 1 (01). [El. Res.]: https://ojs.qarshidu.uz/index.php/mp/article/view/473.

10. Rashidov A., Axatov A., & Nazarov F. ICHKI TAQSIMLASH MEXANIZMIDA MA'LUMOTLAR OQIMLARINI BOSHQARISH ALGORITMI. 2024. DOI: https://doi.org/10.5281/ZENODO.11475559.

11. Ahatov A. R., Ashvini R., Rashidov A. E. Ugli, Nazarov F. M. Optimizaciya kolichestva baz dannyh pri obrabotke bol'shih dannyh. 2023. DOI: https://doi.org/10.24412/2073-0667-2023-1-33-47.

12. Akhatov A. & Rashidov A. Big Data va unig turli sohalardagi tadbiqi // Descendants of Muhammad Al-Khwarizmi, 2021. № 4 (18), P. 135–44.

13. Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., & Cho, Y.-I. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms // Mathematics, 2024. N 12 (16), P. 2553. DOI: https://doi.org/10.3390/math12162553.

14. Rashidov, A., Akhatov, A., & Nazarov, F. The Same Size Distribution of Data Based on Unsupervised Clustering Algorithms. In Z. Hu, Q. Zhang, & M. He (Eds.) // Advances in Artificial Systems for Logistics Engineering III. 2023. Vol. 180, P. 437–447. Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-36115-9_40.

15. Hancock, J. T., & Khoshgoftaar, T. M. Survey on categorical data for neural networks // Journal of Big Data, 2020. N 7 (1), P. 28. DOI: https://doi.org/10.1186/s40537-020-00305-w.

16. Larionov, M. Sampling Techniques in Bayesian Target Encoding (Version 2). arXiv. 2020. DOI: https://doi.org/10.48550/ARXIV.2006.01317.

17. Potdar, K., S., T., & D., C. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers // International Journal of Computer Applications, 2017. N 175 (4), P. 7–9. DOI: https://doi.org/10.5120/ijca2017915495.

18. Sun, W., Cai, Y., & Liu, Y. MSR14 Comparisons of Encoding Techniques for Categorical Features in Linear Regression Models // Value in Health, 2022. N 25 (7), S520. DOI: https://doi.org/10.1016/j.jval.2022.04.1221.

19. Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset // 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. P. 6214–6217. DOI: https://doi.org/10.1109/IEMBS.2009.5334548.

20. Seger, C. An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing. 2018. [El. Res.]: https://api.semanticscholar.org/CorpusID:250534659.

21. Akhatov, A., Rashidov, A., & Renavikar, A. Optimization of the database structure based on Machine Learning algorithms in case of increased data flow. In A. Dagur, K. Singh, P. S. Mehra, & D. K. Shukla, Artificial Intelligence, Blockchain, Computing and Security, 2023. Vol. 2, 1st ed., P. 675–680. CRC Press. DOI: https://doi.org/10.1201/9781032684994-109.

22. Jawthari, M., & Stoffova, V. Effect of Encoding Categorical Data on Student's Academic Performance Using Data Mining Methods. 2020. P. 521–526. DOI: https://doi.org/10.12753/2066-026X-20-068.

ANALYSIS OF CATEGORICAL DATA ENCODING ALGORITHMS

Sh. A. Tursunov, A. E. Rashidov

Samarkand State University named after Sharof Rashidov
703004, Samarkand, Uzbekistan

УДК 519
DOI: 10.24412/2073-0667-2025-2-65-80
EDN: ALXCCT

It is known that the efficiency of artificial intelligence, which is recognized as the most useful tool in all fields, is closely related to several factors. One of these important factors is that the data entering the artificial intelligence algorithms must be in a form that these algorithms understand. That is, since artificial intelligence algorithms are based on mathematical operations and expressions, there must be an opportunity to perform mathematical operators on the incoming data. However, there are projects in which, during the use of artificial intelligence algorithms, data is encountered that does not allow arithmetic operations to be performed. Since discarding this data can negatively affect the result of artificial intelligence, it is necessary to convert it from one form to another. That is, this data is converted to the numeric type. It is known that there are several methods for categorical data encoding, and selecting the best one from these methods is a complex research process. This requires the artificial intelligence user to know not only information about the data set, but also information about all the methods. This research work is aimed at analyzing the methods of categorical data encoding. During the research, 12 different methods of form transformation in text data are studied and analyzed. The advantages and disadvantages of each studied method of encoding are studied. At the same time, a comparative analysis of the studied methods is conducted and a general conclusion is given.

Key words: artificial intelligence, categorical data encoding, data encoding methods.

Introduction. Nowadays, artificial intelligence has become an indispensable and very useful tool for all industries. Data processing using artificial intelligence makes a great contribution to the development of existing industries. Therefore, scientists around the world are conducting research on finding solutions to various problems in many industries using artificial intelligence [1, 2, 3]. During this research work, new algorithms are being developed using certain technologies or based on existing [4, 5, 6]. As a result, the accuracy of finding a solution to a given problem is improving. However, it should be emphasized that the effectiveness of artificial intelligence is closely related not only to the correctness of the algorithm or the correct selection of parameters, but also to the data entering the artificial intelligence [7]. In other words, machine learning requires data with sufficient parameters. These data are not always clean. The data must undergo some initial processing before being fed into artificial intelligence algorithms [8, 9]. These stages include a number of complex processes. One of these processes is the process of converting non-numeric data to numeric data. The main difference between this stage and the other stages is that artificial intelligence algorithms cannot be run without this stage.

© Sh. A. Tursunov, A. E. Rashidov, 2025

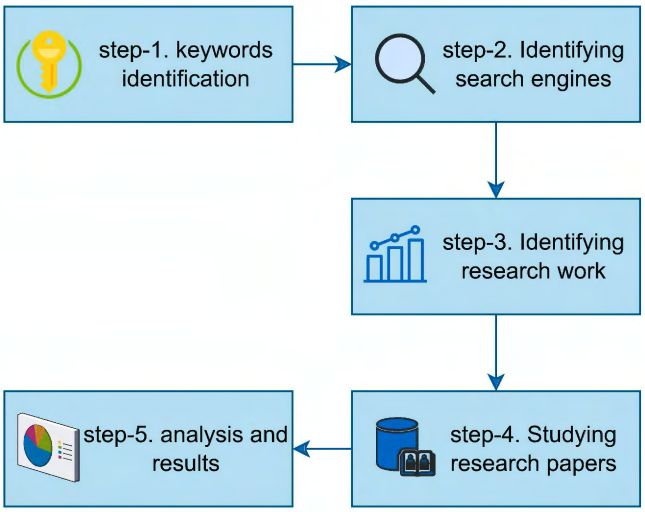

Fig. 1. Research methodology for analyzing data encoding methods in categorical data

Since artificial intelligence algorithms are based on mathematical operations and operators, it is impossible to directly teach text data to artificial intelligence algorithms. Therefore, it is important to first convert them to numeric form.

The process of converting textual data into digital form helps to simplify the data. This makes it easier to perform statistical analysis and machine learning on the data. In addition, digital data usually takes up less memory than textual data, which is effective when working with large volumes of data [10, 11, 12].

Currently, there are several methods for categorical data encoding. However, there is no set of strict rules for their selection or use. Existing studies have suggested choosing an effective method depending on the characteristics of the data. One of the main goals of this research work is to study existing studies and consolidate knowledge about the methods of transforming text data presented in these studies. This study analyzes 12 different methods of data encoding. Popular chatbots widely used in all fields were used to select these methods. The study describes the concept of each algorithm, what data it is effective for, and compares the algorithms with each other through examples. At the same time, a table is presented as a result of the study, which describes the characteristics of the methods of transforming text data.

1. Materials and methods. Methodology. The increasing digital world of today complicates data analysis in any field. This can be seen in the field of scientific research. The study of the analysis of shape-shifting methods in text data is an example of this. It is known that the more methods a study includes, the more data needs to be analyzed. Of course, more data is a good idea, but in this case, problems arise related to the reliability and validity of the data. Therefore, conducting research based on a clear methodology contributes to the effectiveness of the research. For this purpose, this study was conducted based on the methodology presented in Fig. 1.

The research methodology for analyzing the methods of categorical data encoding presented in Fig. 1 includes the following steps:

Step 1: Identifying the most suitable keywords for the research. In this step, keywords were selected to identify the most effective methods among the existing methods. Chatbots such as ChatGPT, Gemini, and Copilot, which are considered effective approaches today, were

used to identify keywords. Based on the requests made to these platforms, attention was paid to considering categorical data encoding algorithms. The responses received from each chatbot were analyzed and the algorithms with a high level of repetition were identified. As a result, a list of 12 different text data shape-shifting algorithms that were considered the most popular and effective was formed.

Step 2: The stage of determining the search engine. The next stage of the analysis and data collection process was the selection of a search engine that would provide effective and comprehensive sources of information. It was decided to use the most popular and widespread search engine — Google — to collect the resources used in this article. The sources obtained through Google served as the basis for ensuring the scientific substantiation of the article and a thorough study of the algorithms.

Step 3: The stage of identifying scientific research materials. At this stage, a search and identification of scientific research on categorical encoding algorithms was carried out. The goal is to identify scientific sources that allow for a complete coverage of modern approaches to algorithms. At the stage of selecting scientific research on shape substitution algorithms, 47 articles were initially selected. Among these articles, 20 articles published over the past 10 years and highly cited were selected.

Step 4: Stage of analysis of scientific research materials. Based on selected scientific studies and articles, a stage of in-depth study of the theoretical and practical aspects of encoding algorithms was carried out. In this process, the application of algorithms to various data, their effectiveness, and their results with various databases, nominal and ordinal data were analyzed. The main methods, approaches, and achievements presented in each article were compared, and the necessary scientific basis for the article was prepared. At the same time, analyses were also conducted to create a more complete picture of the advantages and disadvantages of algorithms.

Step 5: The stage of analysis and results. During the study, the key features of each categorical encoding algorithm were analyzed. In particular, factors such as the ability of the algorithms to work with ordinal and nominal data, the efficiency of working with large amounts of data, and the tendency to overfitting were studied separately. A detailed analysis was conducted of the specific strengths and weaknesses of each algorithm, as well as its suitability for the type of data. At the final stage, a general table was compiled based on the results obtained, indicating the main features of each algorithm. This table serves as a basis for clearly expressing the results of the analysis and comparing different algorithms.

2. Encoding methods for categorical data. One hot encoding. This method creates a separate column for each category of data stored in text columns. Each category is replaced by 0s and 1s. One hot encoding is usually effective when the data does not have any order or value. For example, categories such as “cities”, “occupations”, “countries”, “colors” do not have any value relative to each other. This algorithm is also not considered an effective method for working with large data sets, because the number of columns increases, which can significantly affect the speed of the algorithm by increasing the size of the data [13, 14].

Label encoding. The basic idea of label encoding is to assign a unique integer value to each category in a text variable. For example, if we have a text column called “color” and it has categories like “red”, “green”, “blue”, we can encode them with the numbers 0, 1 and 2, respectively [13, 15].

Label encoding is usually effective when there is a natural order among textual data. Label encoding can change form while preserving the natural order. If there is no natural order among categories, assigning them numerical values using Label encoding can be misleading. Example:

fruits	categorical value of fruits	price
grapes	1	5
apricot	2	10
grapes	1	15
peach	3	20

grapes	apricot	peach	price
1	0	0	5
0	1	0	10
1	0	0	15
0	0	1	20

Fig. 2. Categorical data encoding using one-hot encoding

fruits	type
apple	fruit
orange	fruit
banana	fruit
carrot	vegetable
Tomato	vegetable
Potato	vegetable

fruits	type	type_encoded
apple	fruit	0
orange	fruit	0
banana	fruit	0
carrot	vegetable	1
Tomato	vegetable	1
Potato	vegetable	1

Fig. 3. Categorical data encoding using label encoding

If we have a column called “education”, its categories can be: “secondary”, “higher education”, “incomplete higher education”, which have a natural relationship. Weather: “Rainy”, “Cloudy”, “Sunny” — usually, Rainy weather is the worst, and Sunny is the best.

Target encoding. This is an algorithm used when dealing with text columns with a high cardinality, i.e., a large number of unique categories. For each category, an average target value is calculated and the value is used to shape-shift text data [16].

Target encoding is usually effective when there are a lot of textual data and categories. If there are hundreds or thousands of categories for a single column, algorithms such as one-hot encoding will result in a huge amount of data, which in turn will slow down the algorithm. Target encoding converts textual data into numeric form without adding additional columns, which prevents the data set from becoming too large and sparse. However, such encoding can also cause a number of problems. NaN values pose a risk of overfitting in Target encoding, especially when the data size is small. Techniques such as smoothing or cross-validation are used to prevent this.

Ordinal encoding. Ordinal encoding is a method used to convert categorical variables to numeric format by assigning a unique value to each of their categories. It is also called Label encoding. Ordinal encoding is used when the categories in a variable have a natural order. In this method, the categories are assigned numerical values such as 1, 2, 3, according to their order. For example, if a variable has categories of “low”, “medium”, and “high”, they can be assigned the values 1, 2, and 3, respectively. Ordinal encoding assumes that textual data has a

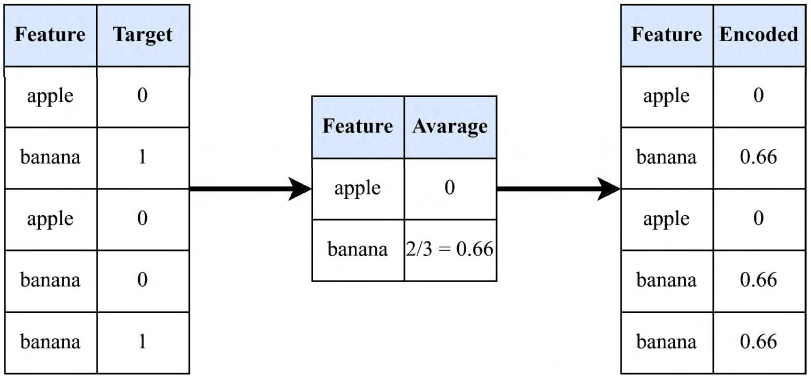

Fig. 4. Categorical data encoding using target encoding

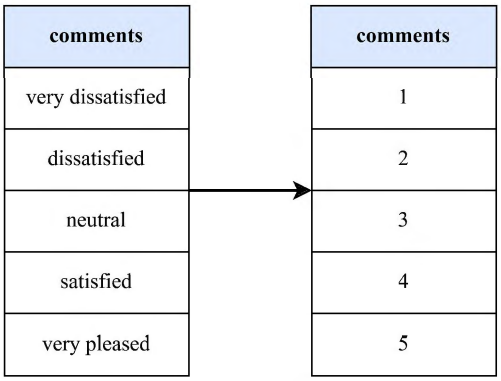

Fig. 5. Categorical data encoding using ordinal encoding.

natural order. It does not work for nominal textual data, since there is no relationship between the categories of a nominal variable [13, 17].

For example, we have a database of customer feedback collected through a survey or online feedback system. It has categories such as “very dissatisfied”, “dissatisfied”, “neutral”, “satisfied” and “very satisfied”. To encode this variable using ordinal encoding, we can assign numeric values as follows:

very dissatisfied — 1, dissatisfied — 2, neutral — 3, satisfied — 4, very satisfied — 5

We code the database by assigning the lowest numerical value 1 to the “very dissatisfied” category and the highest value 5 to the “very satisfied” category. In this way, we achieve the preservation of the natural order of the categories, “very dissatisfied < dissatisfied < neutral < satisfied < very satisfied” $1 < 2 < 3 < 4 < 5$. Suppose we have another text column containing the categories red, blue, and green. We could encode this variable using ordinal encoding by assigning 1 to red, 2 to blue, and 3 to green, but this could lead to incorrect results. This is because there is a natural order between the encoded values - $1 < 2 < 3$, but there is no relationship between red < blue < green.

Dummy encoding. Dummy encoding, like one-hot encoding, uses 0s and 1s for each category in a column to convert text data to numeric form, i. e., it adds new columns for each category. The main difference with one-hot encoding is that dummy encoding uses N-1 features to encode N categories, while one-hot encoding uses N. Dummy encoding discards the first column and changes the form there. The main advantage of dummy encoding is that it adds one less

	One-hot encoding			*Dummy encoding*	
Animals	**cat**	**dog**	**horse**	**dog**	**horse**
dog	0	1	0	1	0
cat	1	0	0	0	0
horse	0	0	1	0	1
dog	0	1	0	1	0
cat	1	0	0	0	0

Fig. 6. Categorical data encoding using one-hot encoding and dummy encoding and their difference

fruits	**frequency encoded**
apple	50
banana	30
cherry	20

Fig. 7. Categorical data encoding using frequency encoding

column, which prevents the algorithm from running slow due to sparse data and data size. Dummy encoding is usually effective when working with small data [13, 16].

We look at a small database of animals, it has 3 categories: “cat”, “dog”, “horse”. We encode this text column using One-hot encoding and Dummy encoding, so we can see the difference between the two algorithms:

Dummy encoding can easily distinguish a row containing a cat even if it has one less column, where two 0s mean it is a cat.

Frequency encoding is another effective method for encoding. It replaces each text string with its frequency, that is, the number of times it occurs in a column [18, 19]. This method is effective when the frequency of text data depends on our result. For example, if we have information about cars and their manufacturers, large manufacturers produce many cars, and small ones much less, this information can be potentially useful. However, there are several disadvantages to this algorithm. If two categories have the same frequency, then after changing the form, they will remain the same frequency, which can lead to the loss of valuable information. For example, we have a text column called Fruits and it contains three fruits — apple, banana, cherry. If apple occurs 50 times, banana 30 times, cherry 20 times. We will convert them to numeric values by replacing them with 50, 30, 20 respectively.

Frequency encoding is very similar to Count encoding, the main difference is that Count encoding records the number of repetitions, while Frequency encoding replaces the term Frequency with a form. Count encoding is a method of encoding text data in a database by counting the number of occurrences [19].

Leave-one-out encoding is a categorical data encoding method that calculates the average value of a column for each category that is associated with all rows except for itself. This average

fruits
apple
banana
apple
banana
banana

→

fruits	count encoded
apple	2
banana	3
apple	2
banana	3
banana	3

Fig. 8. Categorical data encoding using count encoding

city	house_type	price
tashkent	apartment	300
samarkand	house	200
bukhara	apartment	250
tashkent	house	310
samarkand	apartment	210
bukhara	house	240
tashkent	apartment	305
samarkand	house	215
bukhara	apartment	245

→

city	house_type	price	city encoded	house_type encoded
tashkent	apartmen	300	307.5	252.5
samarkand	house	200	212.5	255
bukhara	apartment	250	242.5	265
tashkent	house	310	302.5	218.333
samarkand	apartment	210	207.5	275
bukhara	house	240	247.5	241.667
tashkent	apartment	305	305	251.25
samarkand	house	215	205	250
bukhara	apartment	245	245	265

Fig. 9. Categorical data encoding using leave-one-out algorithm

value is used to convert textual data to numeric form. However, unlike simple encoding, it preserves the meaning of the data [4, 18]. Leave-one-out encoding is effective if text columns (e. g., city, product type) are directly related to related values (e. g., price, sales). This algorithm is useful when you want to avoid overfitting and when the database has high cardinality.

Leave-one-out encoding calculates the average value of the data based on the following formula:

$$leave\ one\ out\ = \frac{(sum\,(values) - \ value)}{(n() - \ 1)}.$$

Let's calculate one column using a formula, for example, the category "Tashkent": there are three Tashkents in the database and their corresponding values: 300, 310, 305. Sum(values) = 300+310+305 = 915, value=300, n=3. *Leave one out* = 307,5.

Binary encoding is similar to one-hot encoding, but instead of creating a separate column for each category, it replaces the category with binary numbers [15, 20]. For example, if a variable has four categories, 'A', 'B', 'C', and 'D', they can be represented as 0001, 0010, 0100, and 1000, respectively.

Binary encoding reduces the size compared to One-Hot Encoding, especially for features with many unique categories. It saves memory, is easy to use, and is simple.

city
samarkand
tashkent
bukhara
namangan

→

city	city encoded
samarkand	00
tashkent	01
bukhara	10
namangan	11

Fig. 10. Categorical data encoding using binary encoding

colors	red	blue
red	1	0
blue	0	1
green	-1	-1

Fig. 11. Categorical data encoding using effect encoding

fruits
apple
fig
apricot
grapes
banana

→

fruits	encoded
apple	01
fig	02
apricot	10
grapes	11
banana	20

Fig. 12. Categorical data encoding using Base N algorithm

Effect encoding, also known as Deviation encoding, Sum encoding, is similar to Dummy encoding, but the main difference is that instead of binary values (0 and 1), three values are used: 0, 1, and -1. This encoding method is particularly useful when working with categorical variables in linear models, as it helps to more effectively address the problem of multicollinearity [21, 13, 22]. For-example we have a text column called “color” and the following three values: red, blue, green.

Before converting the data to numbers, these categories need to be reshaped according to their degree of correlation. This represents how much each category differs from its mean. The base category (green) is separated from the mean of the other two categories. This method is important in interpreting regression results because the coefficients indicate how strong the category effect is relative to the mean.

Base N encoding. In binary encoding, we used the binary number system to encode data, while Base N allows us to encode text data using any number of base numbers. For example, we have a column called “fuits”.

In our example, we converted text data to numeric form using Base-3.

Gray encoding creates a binary-based code for text data, but it is structured in such a way that the values encoding adjacent lines differ by only one bit. This allows for more precise

category	binary encoding	gray encoding
apple	0	0
banana	1	1
peach	10	11
pomegranate	11	10
apricot	100	110
grapes	101	111
cherry	110	101
date	111	100
melon	1000	1100
watermelon	1001	1101

Fig. 13. Categorical data encoding using gray algorithms and difference from binary encoding

transitions between data and reduces the possibility of errors in the machine learning process. For example, there is a column called “fruits”. Let’s see how it works by encoding text data using Binary and Gray encoding:

Imagine you have a sensor that measures 10 different categories (Apple, Banana, etc.) in sequence. If you use a simple binary code and go from “Pomegranate” (0000) to “Apricot” (1001), the code changes 3 bits at a time. But if you use Gray encoding, only one bit changes for each successive category change, meaning the change is smoother and more consistent.

3. Results. The table provides a comparative analysis of various categorical encoding algorithms based on their key features. One-hot encoding and dummy encoding significantly increase the number of categories, leading to high dimensionality, while other methods like label encoding, ordinal encoding, and target encoding are more compact. Target encoding, frequency encoding, and leave-one-out encoding are better suited for handling missing values, whereas label and ordinal encoding preserve inter-category relationships. Binary and Base-N encoding offer dimensionality reduction but lack compatibility with missing data. Overall, each algorithm has distinct strengths, making them suitable for specific use cases depending on the dataset’s structure and model requirements. It should be emphasized that no single encoding technique is universally optimal for all data types. For instance, One-hot encoding yields robust performance on nominal (non-ordinal) textual features by treating each category independently, but its efficiency deteriorates sharply when the number of unique categories is large, leading to prohibitively high memory use and slower training. Conversely, Ordinal encoding is ideally suited for genuinely ordered data—where numerical ranks reflect meaningful progression—but when applied to unordered categories it imposes an artificial hierarchy, potentially misleading tree-based splits or regression coefficients. By explicitly discussing these failure modes alongside the corresponding entries in Table, it provide practitioners with clearer guidance on matching each encoding method to the true nature of their dataset.

Conclusion. The article provides a comprehensive analysis of various methods for converting text data into digital form for artificial intelligence systems. During the study,

Table 1

Analysis of Categorical Encoding Algorithms' Features

A. Increase in dimensionality — Expands the number of features by creating additional columns for categories. B. Preservation of relationships — Maintains the natural relationships between categorical values. C. High dimensionality risk — May cause an excessive number of dimensions, increasing model complexity and computation cost. D. Numerical data compatibility — Ensures that encoded values are in a numerical format suitable for machine learning models. E. Model compatibility — Provides encoded data that can be easily used by a wide range of machine learning algorithms. F. Handling missing values — Offers mechanisms to appropriately manage missing or null category values during encoding. G. Overfitting tendency — May introduce a risk of overfitting, especially when encoding rare categories or small datasets.

Algorithms	A. Increases the number of categories	B. Preserves inter-category relationships	C. Leads to high dimensionality	D. Flexibility with numerical data	E. Compatibility with models	F. Handles missing values	G. Overfitting tendency
One-hot encoding	+		+	+	+		
Label encoding		+		+			+
Target encoding		+		+	+	+	+
Ordinal encoding		+		+			+
Dummy encoding	+		+	+	+		
Frequency encoding		+		+	+	+	+
Count encoding		+		+	+	+	+
Leave-one-out encoding		+		+	+	+	+
Binary encoding				+	+		
Effect encoding		+	+	+	+		
Base-N encoding				+	+		
Gray encoding		+		+	+		

12 different encoding algorithms were studied, their operating principles, strengths and weaknesses were studied and highlighted. The results show that each encoding method has its own characteristics and should be used depending on the nature of the data and the purpose of the study. For example, while One-hot encoding can be problematic when working with large amounts of data, frequency-based methods are effective in solving the volume problem and One-hot encoding method is not recommended for large amounts of data due to the high dimensionality of the resulting matrices, which leads to increased processing time and memory consumption. The summarized table of the algorithms presented can be used to select the most appropriate approach in different situations.

References

1. Nazarov, F. M., O'G'Li, Y. S. S., & O'G'Li, E. B. S. Algorithms To Increase Data Reliability In Video Transcription // 2022 IEEE 16th International Conference on Application of Information and

Communication Technologies (AICT), 2022. 1–6. DOI: https://doi.org/10.1109/AICT55583.2022.10013558.

2. Rashidov, A., Akhatov, A., Aminov, I., Mardonov, D., & Dagur, A. Distribution of data flows in distributed systems using hierarchical clustering. In A. Dagur, D. K. Shukla, N. F. Makhmadiyarovich, A. A. Rustamovich, & J. J. Sindorovich, Artificial Intelligence and Information Technologies. 2024. 1st ed., P. 207–212. CRC Press. DOI: https://doi.org/10.1201/9781032700502-34.

3. Zaynidinov, H., Xuramov, L., & Khodjaeva, D. Intelligent algorithms of digital processing of biomedical images in wavelet methods. In A. Dagur, K. Singh, P. S. Mehra, & D. K. Shukla, Artificial Intelligence, Blockchain, Computing and Security. 2023. Vol. 2. 1st ed., P. 648–653. CRC Press. DOI: https://doi.org/10.1201/9781032684994-105.

4. Nazarov, F. M., Yarmatov, S., & Xamidov, M. Machine Learning Price Prediction on Green Building Prices // 2024 International Russian Smart Industry Conference (SmartIndustryCon), 2024. P. 906–911. DOI: https://doi.org/10.1109/SmartIndustryCon61328.2024.10515790.

5. Rashidov, A., Akhatov, A., & Nazarov, F. The Same Size Distribution of Data Based on Unsupervised Clustering Algorithms. In Z. Hu, Q. Zhang, & M. He (Eds.), Advances in Artificial Systems for Logistics Engineering III. 2023. Vol. 180, P. 437–447. Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-36115-9_40.

6. Rustamovich, A. A., Mekhriddin, N., Fayzullo, N., & Sabharwal, M. Intelligent system of labor market regulation based on the evolutionary modeling of employment // 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2022. P. 2534–2539. DOI: https://doi.org/10.1109/ICAC3N56670.2022.10074149.

7. Rashidov, A., Akhatov, A., & Mardonov, D. The Distribution Algorithm of Data Flows Based on the BIRCH Clustering in the Internal Distribution Mechanism // 2024 International Russian Smart Industry Conference (SmartIndustryCon), 2024. P. 923–927. DOI: https://doi.org/10.1109/SmartIndustryCon61328.2024.10516193.

8. Rashidov Akbar Ergash o'g'li, S. J. S. o'g'li. Selecting Methods Of Significant Data From Gathered Datasets For Research. 2024. DOI: https://doi.org/10.5281/ZENODO.10781255.

9. Rashidov, A., & Madaminjonov, A. Sun'iy intellekt modelini qurishda ma'lumotlarni tozalash bosqichi tahlili: Sun'iy intellekt modelini qurishda ma'lumotlarni tozalash bosqichi tahlili // Modern problems and prospects of applied mathematics, 2024. N 1 (01). [El. Res.]: https://ojs.qarshidu.uz/index.php/mp/article/view/473.

10. Rashidov A., Axatov A., & Nazarov F. ICHKI TAQSIMLASH MEXANIZMIDA MA'LUMOTLAR OQIMLARINI BOSHQARISH ALGORITMI. 2024. DOI: https://doi.org/10.5281/ZENODO.11475559.

11. Ахатов А. Р., Ашвини Р., Рашидов А. Э. Угли, & Назаров Ф. М. Оптимизация количества баз данных при обработке больших данных. 2023. DOI: https://doi.org/10.24412/2073-0667-2023-1-33-47.

12. Akhatov A. & Rashidov A. Big Data va unig turli sohalardagi tadbiqi // Descendants of Muhammad Al-Khwarizmi, 2021. № 4 (18), P. 135–44.

13. Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., & Cho, Y.-I. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms // Mathematics, 2024. N 12 (16), P. 2553. DOI: https://doi.org/10.3390/math12162553.

14. Rashidov, A., Akhatov, A., & Nazarov, F. The Same Size Distribution of Data Based on Unsupervised Clustering Algorithms. In Z. Hu, Q. Zhang, & M. He (Eds.) // Advances in Artificial Systems for Logistics Engineering III. 2023. Vol. 180, P. 437–447. Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-36115-9_40.

15. Hancock, J. T., & Khoshgoftaar, T. M. Survey on categorical data for neural networks // Journal of Big Data, 2020. N 7 (1), P. 28. DOI: https://doi.org/10.1186/s40537-020-00305-w.

16. Larionov, M. Sampling Techniques in Bayesian Target Encoding (Version 2). arXiv. 2020. DOI: https://doi.org/10.48550/ARXIV.2006.01317.

17. Potdar, K., S., T., & D., C. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers // International Journal of Computer Applications, 2017. N 175 (4), P. 7–9. DOI: https://doi.org/10.5120/ijca2017915495.

18. Sun, W., Cai, Y., & Liu, Y. MSR14 Comparisons of Encoding Techniques for Categorical Features in Linear Regression Models // Value in Health, 2022. N 25 (7), S520. DOI: https://doi.org/10.1016/j.jval.2022.04.1221.

19. Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset // 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. P. 6214–6217. DOI: https://doi.org/10.1109/IEMBS.2009.5334548.

20. Seger, C. An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing. 2018. [El. Res.]: https://api.semanticscholar.org/CorpusID:250534659.

21. Akhatov, A., Rashidov, A., & Renavikar, A. Optimization of the database structure based on Machine Learning algorithms in case of increased data flow. In A. Dagur, K. Singh, P. S. Mehra, & D. K. Shukla, Artificial Intelligence, Blockchain, Computing and Security, 2023. Vol. 2, 1st ed., P. 675–680. CRC Press. DOI: https://doi.org/10.1201/9781032684994-109.

22. Jawthari, M., & Stoffova, V. Effect of Encoding Categorical Data on Student's Academic Performance Using Data Mining Methods. 2020. P. 521–526. DOI: https://doi.org/10.12753/2066-026X-20-068.

Турсунов Шерзод Абдувакил угли — магистрант 1 курса факультета интеллектуальных систем и компьютерных технологий Самаркандского государственного университета имени Шарофа Рашидова. Основные научные интересы: искусственный интеллект, кодирование данных, обработка информации. E-mail: sherzodtursunov943@gmail.com

Tursunov Sherzod Abduvakil o'g'li — is a first-year master's student at the Faculty of Intellectual Systems and Computer Technologies, Samarkand State University named after Sharof Rashidov. His research interests include artificial intelligence, data encoding, and information processing. E-mail: sherzodtursunov943@gmail.com

Рашидов Акбар Эргаш угли — заведующий кафедрой искусственного интеллекта и информационных систем Самаркандского государственного университета им. Ш. Рашидова, доктор философии. Область интересов: искусственный интеллект, большие данные, распределенные вычислительные технологии, анализ данных, интеллектуальный анализ данных. Автор более 50 научных статей.

Rashidov Akbar Ergash o'g'li is the Head of the Department of Artificial Intelligence and Information Systems of Samarkand State University named after Sh. Rashidov, PhD. His field of interests includes Artificial Intelligence, Big Data, distributed computing technologies, data analysis, and data mining. Author of more than 50 scientific articles.

Scopus: https://www.scopus.com/authid/detail.uri?authorId=57463410100.

Google Scholar: https://scholar.google.com/citations?user=_zl-wjYAAAAJ&hl=ru.

E-mail: researcher.are@gmail.com

Дата поступления — 15.01.2024