Generate Explanations for Time-series classification by ChatGPT

Zhechang Xue[1,*,†], Yiran Huang[2,†], Hongnan Ma[3,†] and Michael Beigl[4,†]

[1]Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany
[2]Telecooperation Office, Karlsruhe Institute of Technology, Vincenz-Prießnitz-Straße 1, 76131 Karlsruhe, Germany
[3]School of Computer Science, Merchant Venturers Building, 75 Woodland Road, BS8 1TH Bristol, UK
[4]Telecooperation Office, Karlsruhe Institute of Technology, Vincenz-Prießnitz-Straße 1, 76131 Karlsruhe, Germany

Abstract

With the development of machine learning, the concept of explainability has gained increasing significance. It plays a crucial role in instilling trust among clients regarding the results generated by AI systems. Traditionally, researchers have relied on feature importance to explain why AI produces certain outcomes. However, this method has limitations. Despite the existence of documents that introduce various samples and describe formulas, comprehending the implicit meaning of these features remains challenging. As a result, establishing a clear and understandable connection between features and data can be a daunting task. In this paper, we aim to introduce a novel method for explaining time-series classification, leveraging the capabilities of ChatGPT to enhance the interpretability of results and foster a deeper understanding of feature contributions within time-series data.

Keywords

Time-series classification, ChatGPT, Explainability

1. Introduction

With the explosion of AI technology in the last decade, the demand for explainability has increased significantly. However, most models suffer from a lack of explanability. Explainability means the output of machine learning can be interpreted or understood in a clear and easily comprehensible way. This defect can be highly crippling as many domains involving critical decision-making (finance, medicine, etc.) need a model that can explain its decisions to human users.

Classification is one of the most common tasks of the time-series data process. FIR (Feature importance ranking) is one of the most powerful tools to help people understand why models make such decisions. It aims to measure the contributions of individual input features (variables) to the performance of a model [1]. Due to the frequent absence of variables with direct real-world significance in time-series data compared to other data types, extraction of time series

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

jeffxue992@gmail.com (Z. Xue); yhuang@teco.edu (Y. Huang); ex20249@bristol.ac.uk (H. Ma); michael@teco.edu (M. Beigl)
https://lab992.github.io/ (Z. Xue); https://huangyiran.github.io/yiranhuang.github.io/ (Y. Huang)
0009-0002-6458-4749 (Z. Xue); 0000-0003-3805-1375 (Y. Huang); 0009-0004-5767-4089 (H. Ma); 0000-0001-5009-2327 (M. Beigl)

feature becomes one of the most essential preliminary steps [2]. Normally, these features are calculated by complex formulas.

However, these features are not easily understandable. Take *fft_coefficient__attr_'abs'__coeff_30* as an example. This feature is generated by TSFRESH[1], one of the most popular feature extraction packages. The document of TSFRESH, which explain the meanings of every feature, says that this feature means the fourier coefficients of the one-dimensional discrete fourier transform for real input. However, this explanation is purely mathematical and does not easily translate into a physical interpretation in the real world.

An ideal explanation should be: "This data is classified to class [shake hand], since the feature *fft_coefficient__attr_'abs'__coeff_30* has a high value, indicating the presence of a periodic pattern likely caused by a shaking motion." This explanation explicitly links the mathematical definition of the feature to a real-world motion, making it more intuitively understandable.

Given ChatGPT's ability for causal inference, we aim to utilize this ability to connect classification outcomes to easily understandable explanations. Long et al. [3] showcased that LLMs can accurately generate Directed Acyclic Graphs (DAGs) with proper prompts, indicating their capacity for causal reasoning. Moreover, ChatGPT has demonstrated prowess in data analysis. Salim et al. [4] conducted the first time-series forecasting with a prompt-based method, and they concluded that the performance of prediction is better than the numerical-based method.

Based on the previous works, we have found that it's feasible to use LLM to analyze time-series datasets. By far, all the experiments based on LLM are time-series predictions. There's no research about time-series classification and its explanability. Thus, we want to conduct innovative research on this topic.

In this paper, we will explore a method for generating explanations in time-series classification. Our approach involves using ChatGPT to classify data samples and subsequently explaining the rationale behind its classifications. It is important to note that GPT's numerical processing capabilities are comparatively inferior to numerical methods. Therefore, our goal is to trade off a small degree of accuracy in exchange for the ability to generate explanations. The code of this method is available on GitHub.[2]

2. Related Works

Some traditional methods for explaining time-series classification have been proposed. Senin et al. pioneered interpretable time series classification using SAX and vector space models [5]. This approach enables the ranking of time-series patterns by importance, serving as a precursor to FIR. However, despite its interpretability, the classification performance is not ideal.

Motivated by LIME [6], a famous approach to the explanation of classifier, Guillemé et al. [7] proposed the first agnostic Local Explainer For TIme Series classificaTion (LEFTIST), which provides explanations for predictions made by any time series classifier. They showed that explanations generated by LEFTIST can help users understand the classification in easy cases.

Coincidentally, Torty et al. [8] proposed a framework called LIMESegment, which has demonstrated the production of more faithful and robust explanations compared to the existing

[1]https://tsfresh.readthedocs.io/en/latest/

[2]https://github.com/lab992/Generate_explanations_for_classification_by_ChatGPT

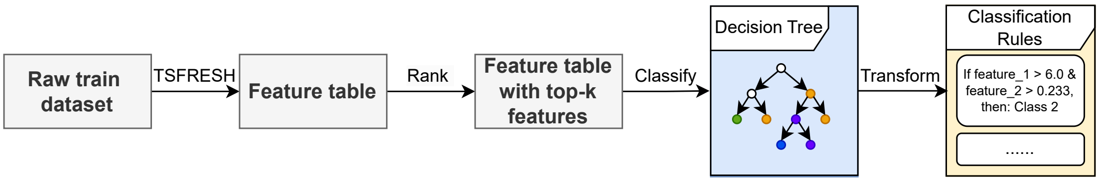

Figure 1: Extract classification rules from train dataset

state-of-the-art adaptation of LIME for time-series classification. (e.g. Neves et al. [9])

We have found that local explanation plays a big role in time-series classification. By observing the pattern of data, reliable explanations can be generated. we can try to combine LLM and local explanation together, to make the explanation not only reliable but also understandable.

3. Methodology

3.1. Components of prompt

The main idea of this methodology is to translate time-series data into a description of the movement. Based on this description, GPT will classify the data into a specific class. Finally, GPT will explain the rationale behind its classification.

To elicit a comprehensive response, a well-structured prompt is essential. A prompt comprises a context and a query. An ideal response should encompass classification and an explanation.

The context component encompasses background information and a description of classification rules. The background part should outline the data measurement scenario, providing ChatGPT with a clear understanding of the mission. This understanding aids ChatGPT in grasping the context and objectives of the subsequent tasks. The description of classification rules acts as a set of instructions derived from the training dataset. These instructions guide ChatGPT in accurately classifying the given description of the test data sample based on features. In the mission part of the query, a precise inquiry should be formulated concerning both the correct classification and an explanation. Additionally, the desired answer format should be specified to enhance readability. It's imperative to instruct ChatGPT to refrain from providing code-based responses and instead use textual explanations.

Subsequently, a detailed methodology for generating descriptions of classification rules and test data samples will be introduced.

3.2. Extract classification rules

Indeed, utilizing data descriptions in the methodology rather than directly inputting raw time-series data into ChatGPT offers several advantages. Firstly, it overcomes the limitation of input length, as raw data often exceeds ChatGPT's input capacity. Secondly, ChatGPT may struggle with understanding multi-digit numbers, which can lead to inaccuracies in analysis and classification [10]. Therefore, employing a method to describe time-series data effectively becomes crucial, and generating classification rules from the training dataset serves as a viable and beneficial approach to address these challenges.

The process of extracting classification rules is outlined in Fig. 1. Absolutely, describing data using features provides a more comprehensive understanding compared to using only numerical values. Features add context and meaning to the data, allowing for a deeper analysis and interpretation of its characteristics. In the initial step, TSFRESH [11] is utilized to construct a feature table comprising hundreds of features from the training dataset. Recognizing that an explanation based on hundreds of features can be perplexing, features generated by TSFRESH will be then ranked using a decision tree and only top-k features can be reserved. Subsequently, the feature table is utilized to train a decision tree model. Finally, classification rules are extracted from the decision tree and translated into a textual form, offering guidance for ChatGPT to accurately classify data.

3.3. Generate lookup table

To ensure the generation of accurate descriptions for both classification rules and test data samples, a lookup table is indeed necessary. This table serves as a reference point, providing the meanings of various features used in the classification rules and test data. With the lookup table in place, the process of translating data into descriptions becomes more precise and consistent, enhancing the overall accuracy of the generated descriptions. This table consists of four columns: 'Feature name', 'Meaning', 'Type', and 'Value'. An example of the lookup table generated by classification rules is presented in Fig. 2. "Feature name" refers to the name assigned to a specific feature. "Meaning" represents the physical interpretation, providing a contextual understanding of its role in the explanation. "Type" distinguishes whether the meaning associated with a feature is a noun or an adjective, aiding in structuring the description appropriately. "Value" quantifies the extent or significance of the meaning associated with a feature, providing additional detail about its impact or amount.

As is shown in Fig. 2, a single classification rule comprises multiple features, many of which are not readily understandable even with references from TSFRESH's documentation. For example, *number_crossing_0* is explained as *the number of crossings*, which still lacks a clear physical interpretation.

To address this issue, we ask ChatGPT to explain these features. The question and ChatGPT's answer are detailed in Fig. 3. ChatGPT answers us that a big value of this feature implies that there are frequent changes in direction or acceleration. This demonstrates that ChatGPT can provide concrete and understandable explanations for features.

As is shown in Fig. 2, the meanings of features are manually selected from the answers provided by ChatGPT.

Lastly, the feature values will also be directly recorded, forming an essential basis for determining how to describe the features.

3.4. Translate data to description

After constructing the lookup table, a translation from data to descriptions can be conducted. The "Type" and "Value" columns in the lookup table play a crucial role in determining how to describe the meaning associated with each feature.

When the count of a feature becomes a crucial consideration, it is defined as a noun.. Phrases

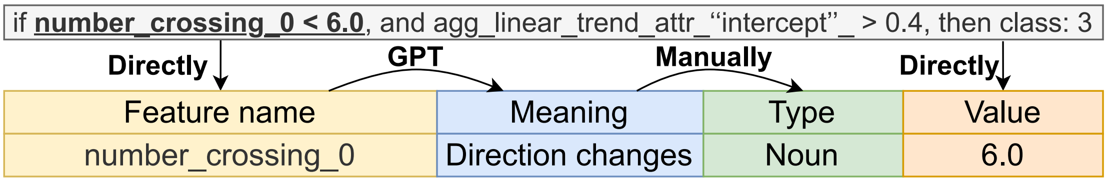

Figure 2: the pipeline of generating lookup table

User: number_crossing_0 is a time series feature, what can you infer from this feature if this value is big during a measurement of accelerometer on x-axis. Give me only one most important pattern.
ChatGPT: A significant value of “number_crossing_0” in the context of accelerometer measurements on the x-axis typically indicates a pattern of rapid and frequent **changes in direction** or acceleration. This could suggest that the object or system being measured is undergoing dynamic and possibly complex movements, such as sudden stops and starts, sharp turns, or oscillations. Such patterns are often associated with activities like vehicular motion, sports activities, or machinery operations where rapid changes in movement direction or intensity occur frequently.

Figure 3: The example of ask GPT to explain the given feature

such as ‘A lot of’ or ‘Few’ are used to describe them. If the provided data feature value exceeds the corresponding value in the lookup table, then the phrase ‘A lot of’ is used; conversely, ‘Few’ is used. Similarly, when emphasizing the extent of the feature’s impact or significance, it is defined as an adjective, terms like ‘Big’ or ‘Slight’ are employed for description,

For classification rules, for instance, consider the rule in Fig. 4 as an example: “if *number_crossing_0* < 6.0, and *agg_linear_trend__attr_“intercept”_* [3] > 0.4, then class: 3 ”. The program automatically finds the corresponding meanings of features in the lookup table and replaces them. The translated rule becomes: “If there are few direction changes and a big decreasing trend, then class 3.”

For a test data sample, an example in Fig. 4 *number_crossing_0* = 9.0 is translated into *A lot of changes in the direction*, since 9.0 exceeds 6.0. Ultimately, the test sample is assembled with the meanings of each feature.

Finally, the background, a description of classification rules, a description of test data samples, and the mission are sent to ChatGPT. ChatGPT will classify the data and provide explanations for its decisions.

4. Evaluation

4.1. Experiment Setup

Dataset Benchmark We use human activity recognition datasets because explanations are more intuitively understandable. For example, we can infer from a sample with a periodic pattern to a periodic activity, which indicates a shaking motion. In this paper, we use 3 datasets:

[3]This feature indicates a decreasing trend.

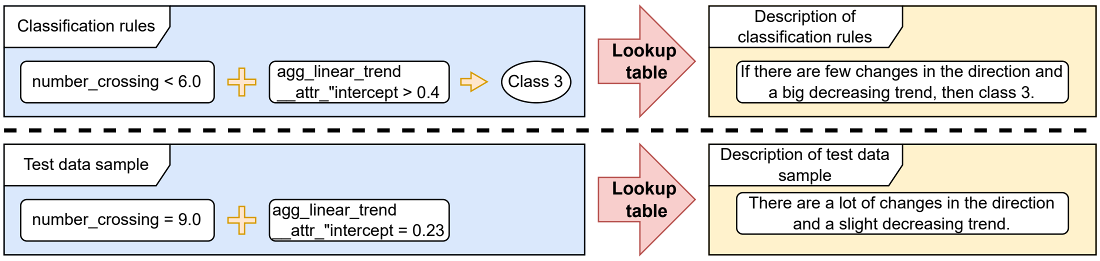

Figure 4: Translate data to description

AllGestureWiimoteX[4], Basketball motion[5], and HMP[6]. For each dataset, we randomly selected three classes.

Evaluate Metrics To assess the performance of our classification models, we employ accuracy as the metric, which evaluates the model's capability to accurately classify the provided samples. Accuracy is determined by the number of correctly classified test data samples divided by the total number of samples.

Models We conducted evaluations using datasets across three models: GPT-3.5-0301, GPT-3.5-0613, and GPT-4-0613.

Baseline In this study, we select a decision-tree as the baseline. We apply the feature selection package TSFRESH to training samples. Then we select the top 3 features of each class and let decision-tree to classify the feature table generated by TSFRESH. The accuracy of AllGestureWiimoteX is 68.33%, Basketball is 66.66%, and HMP is 100%.

4.2. Result

The result is shown in Table 1. Their accuracy is all close to the baseline and therefore it can be concluded that our method is universal for the human activity recognition dataset.

Besides, by observing the accuracy of GPT-models, we conclude that with the advancement of the model, the performance of our method improves, since in all datasets, the experiment on GPT-4 has higher accuracy than on GPT-3.5-turbo-0613, and the experiment on GPT-3.5-turbo-0613 performs better than on GPT-3.5-turbo-0301.

A representative explanation is presented in Fig. 5, where we can see from the explanation that ChatGPT can classify a shaking movement, making the explanation intuitively understandable.

5. Limitation

3-class classification Originally, the three datasets we used in the experiment contained multiple classes. However, given our emphasis on ChatGPT's ability to generate explanations, we opted for a lightweight experiment. As such, we randomly selected three classes from each dataset for analysis.

[4]https://www.timeseriesclassification.com/description.php?Dataset=AllGestureWiimoteX
[5]https://archive.ics.uci.edu/dataset/587/basketball+dataset
[6]https://github.com/wchill/HMP_Dataset

Table 1
The accuracy of experiments on 3 datasets

	Motion	Baseline	GPT-3.5-0301	GPT-3.5-0613	GPT-4
AllGesture	Shake hand	75%	66%	67%	70%
	Move to left	50%	50%	52%	48%
	Move to right	80%	79%	78%	80%
	Total	68.33%	65%	65.67%	66%
Basketball	Dribble	100%	100%	100%	100%
	Hold	100%	100%	100%	100%
	Pass	0%	0%	0%	0%
	Total	66.66%	66.66%	66.66%	66.66%
HMP	Comb hair	100%	24%	80%	92%
	Descend stair	100%	100%	100%	100%
	Lie down bed	100%	100%	100%	100%
	Total	100%	74.67%	93%	97.33%

ChatGPT: According to the given descriptions, your data sample seems to fall into two categories - 'shake hand' due to the lots of changes in direction, and 'move to left' due to a big decreasing trend and few increasing values.
However, the identification of movement is often prioritized by the most distinctive feature that clearly separates one from the others. The 'shake hand' gesture is most prominently characterized by lots of changes in the direction. This aspect, along with the decreasing trend and few increasing values, may suggest that **the hand movement involved turning the hand left and right repeatedly**, while gradually dropping it down a bit - a very likely movement pattern for a shaking hand - thus falling into the 'shake hand' category.
On the other hand, the 'move to left' gesture is characterized by few changes in the direction and a big decreasing trend. Even though these aspects are present in your data sample, the lots of changes in direction do not match the key characteristic of 'move to left', i.e., 'few changes in the direction'. Hence, it would be more accurate to categorize this data sample as 'shake hand'.
So, to summarize: [Class: shake hand].

Figure 5: Explanation of shake hand

Select meaning manually In section. 3.3, the meanings of each feature are manually selected. However, an alternative approach could involve training a model to automatically extract keywords from ChatGPT's answers regarding the meanings of these features.

6. Conclusion and future work

In this paper, we have introduced a method to generate explanations for time-series classification. We have explored the application of LLM in the field of human activity recognition. We have proved that explanations can be generated by our method. By applying this method, the classification performs excellent with a high accuracy. Besides, we have shown that explanations generated by this method are understandable by human evaluation, which means they are close to human cognitive habits.

Furthermore, our pipeline has proved to be universal for other human activity recognition, since all datasets get a result with high accuracy. It's also proved that the performance of our pipeline will be improved with the advancement of models. So far, GPT-4 has proved to be better than GPT-3.5.

Acknowledgment

This work was partially funded by the Ministry of The Ministry of Science, Research and the Arts Baden-Wuerttemberg as part of the SDSC-BW and by the German Ministry for Research as well as by Education as part of SDI-C (Grant 01IS19030A)

References

[1] L. Antwarg, R. M. Miller, B. Shapira, L. Rokach, Explaining anomalies detected by autoencoders using shapley additive explanations, Expert systems with applications 186 (2021) 115736.

[2] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, H. Gamboa, Tsfel: Time series feature extraction library, SoftwareX 11 (2020) 100456.

[3] S. Long, T. Schuster, A. Piché, U. de Montreal, S. Research, et al., Can large language models build causal graphs?, arXiv preprint arXiv:2303.05279 (2023).

[4] H. Xue, F. D. Salim, Promptcast: A new prompt-based learning paradigm for time series forecasting, IEEE Transactions on Knowledge and Data Engineering (2023).

[5] P. Senin, S. Malinchik, Sax-vsm: Interpretable time series classification using sax and vector space model, in: 2013 IEEE 13th international conference on data mining, IEEE, 2013, pp. 1175–1180.

[6] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[7] M. Guillemé, V. Masson, L. Rozé, A. Termier, Agnostic local explanation for time series classification, in: 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI), IEEE, 2019, pp. 432–439.

[8] T. Sivill, P. Flach, Limesegment: Meaningful, realistic time series explanations, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 3418–3433.

[9] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ecgs, Computers in Biology and Medicine 133 (2021) 104393.

[10] J. An, J. Lee, G. Gweon, Does chatgpt comprehend the place value in numbers when solving math word problems, in: Proceedings of the Workshop" Towards the Future of AI-augmented Human Tutoring in Math Learning" co-located with The 24th International Conference on Artificial Intelligence in Education (AIED 2023), Tokyo, Japan, volume 3491, 2023, pp. 49–58.

[11] M. Christ, N. Braun, J. Neuffer, A. W. Kempa-Liehr, Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package), Neurocomputing 307 (2018) 72–77.