ISSN: 2688-8653

Feature Value Allocation of Inspection Data in Cold Chain Logistics with Shapley Value

Ke Liu

School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract

In the process of using cold chain detection data for information mining and analysis, different data features will make different contributions to predictions. To mine the different contributions of these data features, this paper proposes a data feature contribution distribution method based on Shapley value, which is applied to the feature value analysis of the cold chain intelligent sorting prediction system. Through the contribution analysis of the input data features of the Japanese AI "fish face" recognition intelligent sorting project, and comparison with the classic local interpretation method, it is proved that the feature contribution allocation algorithm based on Shapley value is effective.

Keywords

Shapley Value; Cold Chain; Feature Value; Data Value.

1. Introduction

Intelligent decision support for cold chain logistics based on machine learning is reflected in all aspects of the cold chain logistics business process. For the cold chain logistics transportation process, machine learning-based monitoring and early warning can be carried out to ensure the quality of logistics products in real time; for the real-time vehicle distribution path Planning can be intelligently scheduled through machine learning methods; at the same time, the cold chain logistics Internet of Things can be used to predict the shelf life of perishable food and aquatic products. Machine learning and artificial intelligence technologies , as the advanced automated processes, are extremely dependent on huge, real-time data information streams. The cold chain IoT technology can provide the detection data information required by the training model for the cold chain prediction model based on machine learning. The steps of using machine learning algorithms to process the collected data include identifying learning tasks, collecting and cleaning training data, selecting data features for prediction, experimenting with different models and parameters to optimize accuracy, embedding the resulting learning system for prediction and further operations. The machine learning algorithm used in this article is classification algorithm[1,2].

As one of the important methods in cooperative games, the Shapley value method is widely used in the problems of cost allocation and benefit sharing in various industries. Tan et al. [3] used the Shapley value in cooperative game theory to distribute the transmission cost under all loads. Faria et al. [4] used the game theory as a framework to explore the application of different enterprise energy rights allocation methods among hydropower plants. Ma et al. [5] extended the concept of joint game derived from Shapley to encourage selfish Internet service providers to seek to maximize their profits, and eventually converge to the Nash equilibrium. Narayanam et al. [6] studied the diffusion of information in social networks and proposed an influencing node algorithm based on Shapley value. Hou et al. [7] used the linear determination of the Shapley value to determine the allocation based on the service cost-saving game in the case of

ISSN: 2688-8653

cooperative R&D and used the convexity of these service cost-saving strategies to determine the allocation of nuclei, which explained the possibility of innovative enterprise cooperation. Zhang et al. [8] studied a general mobile traffic offloading system, using the widely used Gale-Shapley algorithm to optimize the allocation of mobile phone users (MUs) to offloading stations, and solve the effective bandwidth allocation and charging required between cloud providers and cloud users mechanism. Shi et al. [9] proposed the first dynamic pricing mechanism for on-demand bandwidth between data centers through auctions based on Shapley values and proposed an online traffic scheduling algorithm. Li et al. [10] proposed a service composition method based on the extended Gale-Shapley algorithm, which allows multiple service composition solutions to be efficiently generated. The above-mentioned literature combines the Shapley value method with the different characteristics of the research industry for application or improvement.

The Shapley value research in the logistics industry mostly considers the allocation of allocation costs and the allocation of benefits of cooperative coalitions. Xie et al. [11] used Shapley as the core tool to study the cost-sharing model coalition of cold chain logistics common allocation from the perspective of cooperative games and calculated the marginal contribution rate of each participant. Xu et al. [12]studied the individual profit problem in the fourth-party logistics supply chain coalition system and proposed an improved weighted Shapley value model. Numerical studies have shown that this method fully considers contributions and risks, and effectively distributes profits. There are also related articles that consider the service cost and value distribution of enterprise operation management[13,14].

In the field of data science, it is important to understand why a predictive model makes certain predictions. A more popular method is to interpret the predictions from complex models by estimating the importance of input features. Cohen et al. [15] proposed and studied a feature selection algorithm based on multiple disturbance Shapley analysis, which relies on game theory to evaluate the usefulness, ranks the importance of training data points, and is used to understand model behavior and detect data set errors. Lundberg et al. [16] proposed a unified framework, SHAP (Shapley additive interpretation) to explain predictions. SHAP assigns a specific predicted important value to each feature. Kernel SHAP is a kernel-based Shapley inspired by local proxy models. The value estimation method can approximate the Shapley value under the assumption that all features are independent and uncorrelated. Aas et al. [17] extended Kernel SHAP to solve the related feature allocation problem. Wang et al. [18] studied the symmetric and weighted Shapley values of cooperative n-person games, and designed the weighted Shapley value payout function by defining the asymmetric weights on the participants to ensure that the optimal allocation is a pure Nash equilibrium. Jia et al. [19] used Shapley value to study the data allocation problem, defined a unique payment scheme, and proposed a set of efficient Shapley value approximation algorithm. Lundberg et al. [20] proposed Tree SHAP, an effective estimation method based on a tree-based model, using personalized feature attribution visualization methods to improve the traditional attribution summary and partial dependency graph, and proposed a unique feature attribution-based supervised clustering method . The research on approximate Shapley value is mostly to improve the approximate algorithm from different angles to obtain better results and time complexity.

The current research is mostly focused on applying the Shapley value method to the design of cost allocation or benefit allocation mechanism in various industries but does not consider the use of Shapley value in the cold chain machine learning model of the detection data feature contribution allocation problem. Based on this, the design is based on The fair revenue segmentation algorithm of Shapley value, this is the first time that Shapley values are used to assign characteristic values of cold chain inspection data. Starting from the cold chain aquatic product intelligent sorting model, an approximate Shapley value algorithm is proposed to be applied to the Japanese fish category prediction case In, the contribution value allocation of the

ISSN: 2688-8653

cold chain detection data features input in the machine learning model, and the feature value feedback to improve the prediction model to obtain better prediction results and provide support for cold chain decision-making.

2. Shapley Value in the Machine Learning Algorithm

In the machine learning model training prediction, each feature value of the sample is used as the "player" in the game, and the "game" is the task of predicting a single sample of the data set. The prediction process is "payment"."Profit" is the actual predicted value of the sample minus the average predicted value of all samples, and the Shapley value of the feature is the average marginal contribution of the feature in all feature sequences. The fair allocation method based on Shapley value has the following 4 properties:

Property 1 Effectiveness. $\sum_{i=1}^{n}\varphi_i(V)=V(N)$. That is, all values are allocated, and the sum of the profits of all players participating in the cooperation is all the value generated by the coalition.

Property 2 Symmetry. $\forall i,j\in[N],\forall S\subset[N]\backslash\{i,j\}$, If $V(S\cup\{i\})=V(S\cup\{j\})$, then, $\varphi_i=\varphi_j$. That is, if the contribution of player i and player j are the same, the Shapley value of the two is equal.

Property 3 Redundancy. $\forall i\in[N],\forall S\subset[N]$, If $V(S\cup\{i\})=V(S)$, then $\varphi_i=0$. That is, the Shapley value of uncontributed features is 0.

Property 4 Additivity. $\forall V_1,V_2$, $\varphi(V_1+V_2)=\varphi(V_1)+\varphi(V_2)$. That is if the game player completes two tasks and the income function is V_1,V_2 respectively, then the result of the two tasks' income distribution together should be consistent with the result of the separate distribution.

For the player *i* in the cooperative game, its revenue is distributed as:

$$\varphi_i(N,V)=\sum_{S\subseteq N\backslash\{i\}}\frac{|S|!(|N|-|S|-1)!}{N!}\left(V(S\cup\{i\})-V(S)\right) \tag{1}$$

Where *N* is the coalition formed by all game participants in the cooperation, $N=\{1,2,...n\}$. *S* represents an arbitrary subset of coalitions formed by all game players *N* after removing player *i*, ie. $S\subseteq N\backslash\{i\}$; $V(\cdot)$ is the revenue function, reflects the influence of player i on coalition $S\cup\{i\}$; $\frac{|S|!(|N|-|S|-1)!}{N!}$ is the probability of the player participating in the coalition *S*. The value of each player is $(\varphi_1,\cdots,\varphi_n)$.

The Shapley value method is used in the interpretation of machine learning features because it uniquely possesses the above-mentioned properties required in the concept of data values, such as the fairness and additivity of values in the use of multiple data. The Shapley value is the only solution that satisfies the four properties of symmetry, effectiveness, additivity, and redundancy. At the same time, not only the influence of a single variable is considered, but also the influence of variable groups and the possible synergistic effects between variables are considered. The Shapley value is the only fair allocation scheme that satisfies the four properties and is widely used in various industries. However, the Shapley value has a fatal flaw, that is, its calculation is related to the number of players. The number of players must list all the 2^n types of coalitions. Set and then accurately calculate the Shapley value, which requires a lot of calculation time. In the allocation of the contribution of data information features using

ISSN: 2688-8653

Shapley values, this situation is even more difficult, because the "profit" is the prediction result of machine learning, and the complexity of the model increases exponentially with the increase of feature *n*. In 99.9% of real-world problems, only solutions that approximate the Shapley value are feasible. Therefore, in recent years, scholars have focused on the study of approximate Shapley values, and the "absence" of features is drawn by random examples.

3. Case Analysis

At present, it usually consumes a lot of manpower and material resources in the sorting scene of cold chain aquatic products with a large variety and large quantity. Take the Japanese fishery as an example. To ensure the freshness of the catch, it needs to be sorted, washed, and refrigerated in the shortest time. An experimental project launched in Hachinohe City, Aomori Prefecture, Japan, uses equipment equipped with AI (Artificial Intelligence) systems and cameras to replace manpower to automatically classify a large number of live fish caught. Using salmon, mackerel (i.e. mackerel) , and herring as learning samples, the machine learning system on the device was trained. The training data includes the size, shape, fatness, color, and other data of each fish. The imaging equipment can classify fish photos captured by the camera according to various parameters. The equipment can sort up to 100 fish per minute and can distinguish about 40 kinds of fish, with an accuracy rate of 90%. When the equipment recognizes the type of fish, it will be pushed into the corresponding box, so that each fish passing through the conveyor belt will be accurately identified and sent to the corresponding sub-packing box. The equipment can be used in the transportation of cold chain logistics to carry out fast and effective sorting.

Take the Japanese AI "Fish Face" recognition project as an example to reveal the application of machine learning and artificial intelligence in the prediction of fish species in aquatic products. Combined with the Shapley value model, the global and local contribution analysis of each input feature of species prediction can be performed. Reduce costs and increase efficiency for data-driven logistics sorting systems. Next, the source of the data will be described, then numerical experiments will be carried out, and finally, the experimental results will be explained and analyzed.

3.1. Data Description

Table 1. Part of the original data of the experimental data set

	Species	Length1	Length2	Length3	Height	Width	Weight
1	5	19	20.5	22.8	6.4752	3.3516	0
2	3	7.5	8.4	8.8	2.112	1.408	5.9
3	6	9.3	9.8	10.8	1.7388	1.0476	6.7
4	6	10.1	10.6	11.6	1.7284	1.1484	7
5	6	10	10.5	11.6	1.972	1.16	7.5
6	6	10.8	11.3	12.6	1.9782	1.2852	8.7
7	6	10.4	11	12	2.196	1.38	9.7
8	6	10.7	11.2	12.4	2.0832	1.2772	9.8
9	6	11.4	12	13.2	2.2044	1.1484	9.8
10	6	11.3	11.8	13.1	2.2139	1.1659	9.9

This article uses public data sets from the Kaggle website resource library to conduct experiments to verify the effectiveness of the algorithm. This experimental data set records the data of 7 common fishes. Using this data set, a good machine learning prediction model can be

ISSN: 2688-8653

established and the species of fish can be predicted. The experimental data set is used for classification, with no missing data, multiple data types, and contains 159 data samples. Experimental data set contains 7 related attributes, namely : Species, Length1 (vertical length), Length2 (diagonal length), Length3 (horizontal length), Height, Width (diagonal width) and Weight. Among them, the species of fish are Bream, Parkki, Perch, Pike, Roach, Smelt, Whitefish, a total of seven species (denoted as 0-6). Use the first six attributes to describe the basic characteristics of fish, and the response variable is the type of fish. Part of the original data of the experimental data set is shown in Table 1.

3.2. Experimental Results

The underlying model uses the Xgboost training model and uses the experimental data set to train the model. The global interpretation result is shown in Figure 1.

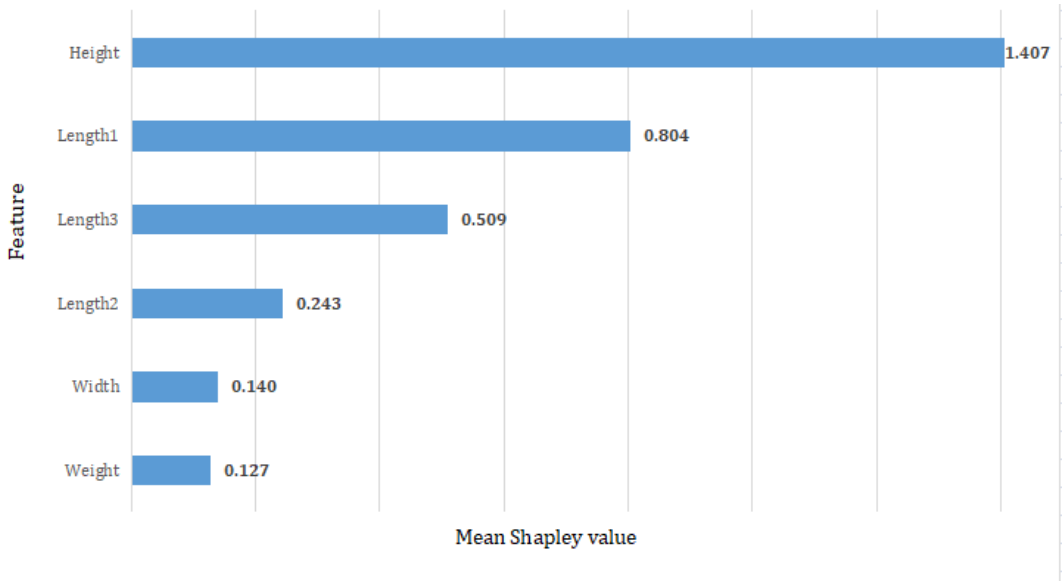

Figure 1. Mean Shapley value of each feature

Figure 1 shows the mean Shapley value of each feature. To illustrate the superiority of applying Shapley value in the interpretation of cold chain detection data contribution, the LIME (Local Interpretable Model-Agnostic Explanations) algorithm is used for comparison. LIME does not require model adaptation. It is a method of interpreting machine learning models that has nothing to do with the model. Therefore, it does not go deep into the model. It only analyzes the input feature values by making small disturbances around it and observing and predicting behavior. The contribution of the input feature to the prediction result. The results of running the experimental data using the LIME method are shown in Figure 2.

In Figure 2, the blue color indicates that the feature has a positive effect on the prediction result of the example, and the red color indicates that the feature has a negative effect on the prediction result of the example. Selecting the LIME partial interpretation of 4 random samples without changing the parameters, it can be seen that the results of each sample are quite different and the reliability is not high. The superiority of the Shapley value is manifested in: First of all, the Shapley value is the only explanation method with a solid theory. A solid theory can give a reasonable explanation of the results, but LIME assumes that the machine learning model is locally linear and has no solid theory to support it. Secondly, due to the validity nature of the Shapley value, it can ensure that the difference between the prediction result and the average prediction is fairly distributed among the feature values of the instance, while LIME cannot guarantee that the prediction is fairly distributed among the features. Third, the Shapley

ISSN: 2688-8653

value allows comparative interpretation. It can not only compare the prediction of a single instance with the average prediction of the entire data set but also compare it with a partial subset or even a single data point. The LIME local model cannot do it, and LIME can only be used for the local interpretation of a single sample. Shapley can not only explain the feature globally but also explain the feature information of a local single sample.

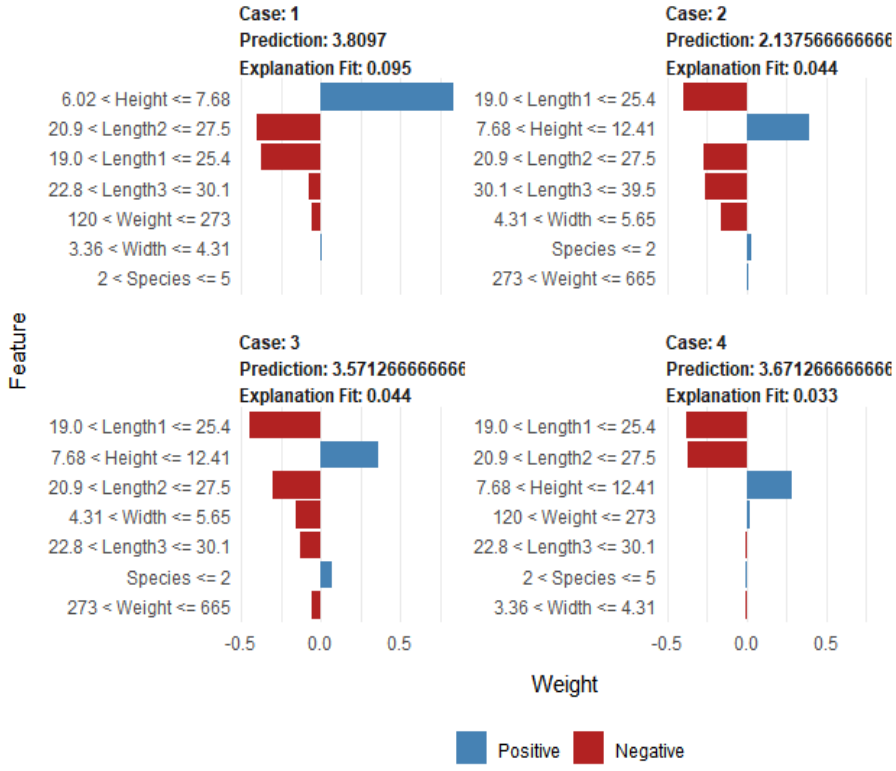

Figure 2. Local interpretation results based on the LIME method

4. Conclusion

This article mainly introduces the construction of an interpretation model for analyzing the value of input features in the predictive sorting of aquatic fish species. Based on the Shapley value fair allocation method of cooperative game, this paper treats prediction as a game process in the machine learning model, treats data features as game parties, cooperates to complete the prediction, and distributes contributions to the generated prediction results. Since the value benefit of machine learning is the prediction result, the number of calculations will increase exponentially in the process of increasing the number of features. This paper designs an approximate Shapley value algorithm for the realization of feature contribution allocation. We take the artificial intelligence sorting of Japanese aquatic fish as an example, analyze the data results, and use the local interpretation model LIME to compare the results. The Shapley value method is used to calculate and analyze the characteristic value of fish prediction results, which provides theoretical results for the development of the cold chain intelligent sorting automation system.

Based on the established model, the use of the Shapley method for feature contribution interpretation can be divided into two levels. At the global level, the Shapley distribution can be used to describe the specific impact, law, and correlation of features; at the local level, the model can give each one the quantitative contribution of each feature in the sample prediction process. After the Shapley algorithm is used to obtain the value contribution of each feature, it can be balanced with the cost of data collection. In the case of artificial intelligence sorting of

ISSN: 2688-8653

Japanese aquatic fish, it is necessary to identify fish-related data such as height, width, etc. To train a model for predicting fish species. Assuming that the prediction cost of collecting 6 features is 1,200 yuan, in large-scale fish species identification, the training model is already "smart" enough and the prediction accuracy is high. Consider reducing the cost of 200 yuan, while still having higher requirements for the prediction results, using the Shapley value method in this article, you can delete the feature with the Shapley value ranking last that contributes the least to the prediction result., i. e. the feature “Weight”. So, the feature is excluded from the collection of input data to save 200 yuan in forecasting costs. Whether in the global interpretation, the Shapley value can get the respective contribution of each feature, or the local single sample, the relevant Shapley value information can be obtained from the prediction result for a specific fish, based on this information, the cold chain detection data prediction task Balance between cost and prediction accuracy, thereby improving input characteristics, optimizing training models, providing effective support for subsequent optimization and improvement of intelligent equipment, providing technical support for automated sorting, and providing result support for cold chain decision-making.

References

[1] Yu H , Shen J , Xu M . Temporal case matching with information value maximization for predicting physiological states[J]. Information ences, 2016, 367:766-782.

[2] Yu H , Shen J , Xu M . Resilient parallel similarity-based reasoning for classifying heterogeneous medical cases in MapReduce[J]. Digital Communications and Networks, 2016, 2(3):145-150.

[3] Tan X, T T Lie. Application of the Shapley Value on transmission cost allocation in the competitive power market environment[J]. IEEE Proceedings - Generation, Transmission and Distribution, 2002, 149 (1):15 -20.

[4] Faria E, Barroso L A, Kelman R, et al. Allocation of Firm-Energy Rights Among Hydro Plants: An Aumann–Shapley Approach[J]. IEEE Transactions on Power Systems, 2009, 24(2):541 -551.

[5] Ma R T B, Chiu D M, Lui J C S, et al. Internet Economics: The Use of Shapley Value for ISP Settlement[J]. IEEE/ACM Transactions on Networking, 2010, 18(3):775 -787.

[6] Narayanam R, Narahari Y. A Shapley Value-Based Approach to Discover Influential Nodes in Social Networks[J]. IEEE Transactions on Automation Science and Engineering, 2011, 8(1):130 -147.

[7] Hou D, Driessen T, Sun H. The Shapley value and the nucleolus of service cost savings games as an application of 1-convexity[J]. IMA Journal of Applied Mathematics, 2015, 80(6):1799 -1807.

[8] Zhang Y, Mao Y, Zhong S. Joint Differentially Private Gale–Shapley Mechanisms for Location Privacy Protection in Mobile Traffic Offloading Systems[J]. IEEE Journal on Selected Areas in Communications, 2016, 34(10):2738 -2749.

[9] Shi W, Wu C, Li Z. A Shapley-Value Mechanism for Bandwidth on Demand between Datacenters[J]. IEEE Transactions on Cloud Computing, 2018, 6(1):19.

[10] Li F, Zhang L, Li Z, Liu Y, et al. QoS-Aware Service Composition in Cloud Manufacturing: A Gale–Shapley Algorithm-Based Approach[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(7):2386 -2397.

[11] Xie R, Pi X, Yang W. Shapley Method for Cost Allocation Research in Cold Chain Logistics Joint Distribution Coalition[J]. DEStech Transactions on Social Science, Education and Human Science, 2017 (emse).

[12] Xu N. Improved weighted Shapley value model for the fourth party logistics supply chain coalition[J]. Journal of Control Science and Engineering, 2013, 2013.

[13] Yu H , Wang Y , Wang J N , et al. Causal Effect of Honorary Titles on Physicians' Service Volumes in Online Health Communities: Retrospective Study[J]. Journal of Medical Internet Research, 2020, 22 (7): e18527.

ISSN: 2688-8653

[14] Yu H Y , Chen J J , Wang J N , et al. Identification of the Differential Effect of City-Level on the Gini Coefficient of Health Service Delivery in Online Health Community[J]. International Journal of Environmental Research and Public Health, 2019, 16(13):2314.

[15] Cohen S, Ruppin E, Dror G. Feature Selection Based on the Shapley Value[C] //IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, 2005.

[16] Lundberg S M, Lee S I. An unexpected unity among methods for interpreting model predictions, arXiv preprint arXiv:1611.07478,2016.

[17] Aas K, Jullum M, Loland, et al. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values[J].2019.

[18] Wang Y, Cheng D, Liu X. Matrix expression of Shapley values and its application to distributed resource allocation[J]. Science China Information Sciences, 2019, 62(2): 1-11.

[19] Jia R, Dao D, Wang B, et al. Towards Efficient Data Valuation Based on the Shapley Value[J]. 2019.

[20] Lundberg S M, Erion G G, Lee S I, Consistent Individualized Feature Attribution for Tree Ensembles, arXiv preprint arXiv: 1802.03888,2018.