Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 4
ICCIA 2024

Validity of data estimation methods in large-scale insurance datasets

Jinrong Zhao[1], Daibin Lan[1], Tengfei Meng[1], Wanting Kan[2]

[1] Guilin university of technology at nanning, Guangxi, Chongzuo, 541006, China

[2] Guangdong University of Finance & Economics, Guangdong, Guangzhou, 510320, China

Abstract. In the insurance industry, accurately processing and analysing large-scale datasets is critical for risk assessment and decision-making. In this study, a comprehensive database was built by collecting and integrating weather-related data, insurance industry data, and attribute-specific data to support in-depth analyses of the impact of extreme weather events. In the data preprocessing stage, we adopted mean-filling and plurality-filling methods to deal with missing data, while applying the 3σ rule to deal with outliers in the data to ensure the completeness and consistency of the dataset. In addition, we used multi-criteria decision analysis methods (VIKOR) and hierarchical clustering models (BRICH algorithm). These advanced analysis techniques not only optimise the data processing process, but also improve the reliability and accuracy of the analysis results. Through these methods, we are able to effectively identify and classify different risk levels, which in turn provides a scientific basis for the pricing of insurance products and risk management. This study shows that advanced data estimation methods can provide effective and accurate support in processing large-scale insurance datasets, which is of great practical significance to the development of the modern insurance industry.

Keywords: Large-scale datasets; Insurance industry; Risk assessment; Data pre-processing; VIKOR methodology.

1. Introduction

In today's insurance industry, the processing and analysis of large-scale data sets is an integral part of the decision-making process. With the advancement of technology, our team has adopted a highly technical approach to data collection, integrating weather-related data, insurance industry data, and attribute-specific data to improve the accuracy and comprehensiveness of our research. Especially in the context of frequent extreme weather events, accurate and comprehensive data analyses are important for assessing the associated risks [1]. However, the completeness and accuracy of data collection is often limited by the problem of missing and outliers in the data itself. To address this issue, we employ mean-population and plurality-population methods to fill in the missing data and deal with outliers through the 3σ principle to ensure the consistency and reliability of the dataset [2]. This preprocessing step is the basis for building an effective decision support system, which helps us to reduce the noise in the data and enhance the interpretive power of the data. In this paper, we will further explore the effectiveness of these data estimation methods on large-scale insurance datasets [3]. Through the application of various data processing techniques, we are not only able to provide more accurate risk assessment, but also data support for insurance product pricing and risk management [4]. The application of this methodology is critical to understanding and responding to complex problems in the insurance domain, especially when dealing with big data environments that involve large amounts of computation and analysis [5]. Through real-world case studies, we will demonstrate the practical effectiveness of data preprocessing techniques in improving the quality of data analyses and the accuracy of decision-making, providing new perspectives and methodological support for data-driven decision-making in the insurance industry.

2. Data description

Our team used a highly technical data collection methodology to acquire data that fell into three main categories: weather-related data, insurance industry data, and property-specific data. Our data collection process included multiple open-source data sources, and the following is a high-level detailed description of our data collection methodology and data sources: weather-related data Historical weather data was collected from multiple trusted weather data providers and government weather agency websites [6]. These datasets include a variety of weather metrics such as temperature, rainfall, humidity, and wind speed, as well as detailed information on extreme weather events such as hurricanes, floods, wildfires, and earthquakes. We ensure the accuracy and completeness of the data in order to provide a reliable weather database for our analyses. Insurance Industry Data A large amount of insurance industry data is crawled from several reliable insurance industry databases and financial data provider websites. These data include historical claims data, total insurance claims, total insurance operating expenses, insurance density, total gross premiums, and penetration rates [7]. These metrics are critical to our research. Property-Specific Data In addition to the above data, we also collect property-specific data, which may include information such as real estate market data, property types, building structures, and historical values as shown in Figure 1.

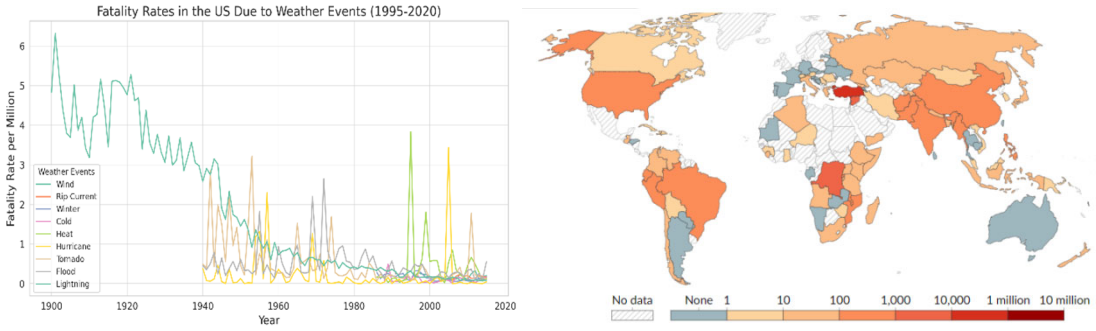

Figure 1. Fatality rates in the US due to weather events; Number of deaths from disasters

These data help us to gain a more comprehensive understanding of property characteristics in different regions in order to better assess the risks associated with extreme weather events.

Through the use of advanced data collection methods and automated tools, we have successfully constructed a complete dataset that provides a solid foundation for our research [8]. The sources of these data were carefully selected to ensure the credibility and accuracy of the data to support our in-depth analysis of the impact of extreme weather events on the insurance industry. The sources of these data can be seen as shown in Table 2 below:

Table 1. Description of data sources

Description	***Website***
Precipitation Data	WebMET - Free Met Data - United States, North Dakota
GIS Data	GeoScene Online
Historical Claims Data	*https://stats.oecd.org/Index.aspx?DatasetCode=INSIND*
Insurance Industry Development	*https://www.swissre.com/institute/research/sigma-research/sigma-2023-03/5-charts-wold-insurance-2023.html*
Disaster Data	*https://ourworldindata.org/natural-disasters?TB_iframe=true&width=370.8&height=658.8*

After constructing the new dataset it was found that there is still a lot of work to be done to preprocess the collected data and fill in the missing data values. For quantitative data features, the data were filled using the mean fill method, and for some special data features, the data were filled using the mode fill method [9]. For the filled data, the outliers are processed using 3sigma outliers. And set it to the mean. After doing this we have a complete data set.

3. VIKOR Model

Similar to the entropy method, the multi-criteria compromise solution ranking method is a multi-criteria decision analysis method. The method was proposed by Serafim Opricovic and Gongzhi Fan in 1998. We use the multi-criteria compromise solution ranking method to transform the decision problem into multiple criteria, establish the weights of each criterion, and obtain the best solution by calculating the degree of superiority of each solution over the others. Its ability to deal with multiple indicators and different types of data has good applicability to the indicators we selected [10]. The specific steps of the multi-criteria compromise solution ranking method are shown in Table 2 below:

Table 2. VIKOR

Standardization of indicators: homogenization of heterogeneous indicators	
Step 1	1.For positive indicators: $r_{ij} = (x_{ij} - x_{ij}^{-})/(x_{ij}^{*} - x_{ij}^{-})$ 2. For negative indicators: $r_{ij} = (x_{ij}^{*} - x_{ij})/(x_{ij}^{*} - x_{ij}^{-})$
Step 2	Establishing group utility versus individual regret Group utility values: $s_i = \sum_{j=1}^{n} w_j\,(b_j^{*} - b_{ij})/(b_j^{*} - b_j^{-})$ Individual regrets: $R_i = \max_{1 \le j \le n}[w_j\,(b_j^{*} - b_{ij})/(b_j^{*} - b_j^{-})]$
Step 3	Calculate the value of the trade-off decision indicator: $Q_i = \frac{v(S_i - S^{*})}{S^{-} - S^{*}} + \frac{(1-v)(R_i - R^{*})}{R^{-} - R^{*}}$ Among them：$S^{*} = min_1 \le i \le mS_i, S^{-} = max_1 \le i \le mS_i$ $R^{*} = min_1 \le i \le mR_i,\ R^{-} = max_1 \le i \le mR_i$

By SPSS software we derive the optimal and worst values of group utility value and individual regret value as shown in Table 3 below.

Table 3. Group utility values and individual regret values

S^{+}	S^{-}	R^{+}	R^{-}	Decision-making mechanism factor v
0.821	1	0.104	0.118	0.5

According to the results of the group utility value and individual regret value, based on which the decision-making indicator Q value is calculated, the smaller the indicator Q value, the better the program, and finally get the ranking.

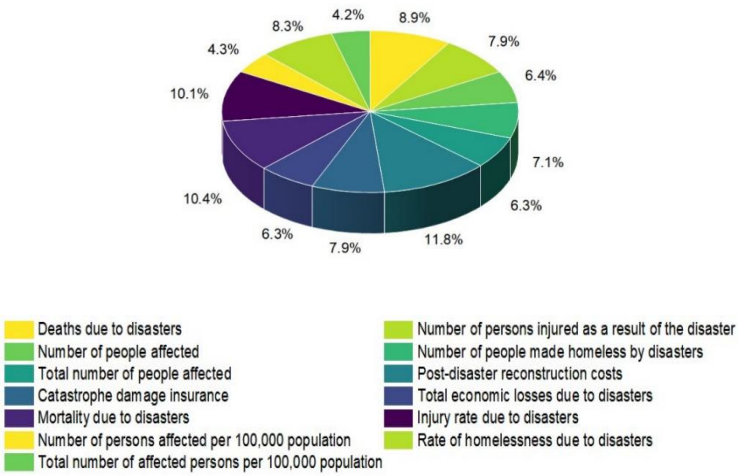

Figure 2. The resulting Q-values of the decision metrics are then classified using a Hierarchical Clustering Model.

4. Hierarchical Clustering Model

In the hierarchical clustering stage of BRICH clustering algorithm, firstly, on the basis of the sub-clusters generated in the initial stage, further hierarchical clustering is performed on each sub-cluster, which is gradually subdivided into smaller sub-clusters. The BRICH algorithm adopts bottom-up hierarchical clustering to merge similar sub-clusters step by step until the stopping condition is satisfied.

Initial sub-clusters: the sub-clusters generated in the initial stage are used as inputs for the initial hierarchical clustering.

Distance or similarity calculation: for each pair of sub-clusters, the distance or similarity between them is calculated. The distance measure here can be the Euclidean distance between the centroids of the subclusters or other suitable distance measures, and the similarity measure can use cosine similarity, correlation coefficient, and so on. We choose Euclidean distance as the distance metric and correlation coefficient as the similarity metric. For the centroids x and y of two subclusters, the Euclidean distance can be expressed as:

$$d(x,y) = \sqrt{(x^1 - y^1)^2 + (x^2 - y^2)^2 + \ldots + (xn - yn)^2} \quad (1)$$

where x_1, x_2, ... , xn and y_1, y_2, ... , yn are the coordinates of the centroids of the two subclusters in each dimension, respectively.

For the feature vectors x and y of the two sub-clusters, the correlation coefficient can be expressed as:

$$r(x,y) = cov(x,y) \,/\, (\sigma x * \sigma y) \quad (2)$$

Where $cov(x,y)$ denotes the covariance between the feature vectors x and y, σx denotes the standard deviation of the feature vector x, and σy denotes the standard deviation of the feature vector y.

Nearest Neighbor Merge: find the two subclusters with the closest distance or similarity and merge them into a new subcluster. Here merging can be done using methods like single linking, complete linking or average linking. We choose single link method for merging, in single link merging method, the distance or similarity $d(Ci, Cj)$ of merging two subclusters Ci and Cj can be calculated by the following formula:

$$d(Ci, Cj) = min\big(d(p,q)\big) \quad (3)$$

where $d(Ci, Cj)$ denotes the distance or similarity between sub-clusters Ci and Cj, and $d(p, q)$ denotes the distance or similarity between data points in sub-cluster Ci p and the distance or similarity between the data point q in subcluster Cj. In the process of hierarchical clustering, we calculate the distance or similarity between all possible subclusters and select the two nearest neighboring subclusters for merging. So, in each merge, we iterate through all the combinations of subclusters Ci and Cj, calculate the distance or similarity between them and select the minimum value as the merged distance or similarity.

Update subclusters: for the newly formed subclusters, update the labeling of the subclusters by relabeling the data points of the two subclusters before the merger as new subcluster labels.

Repeat step 3 and step 4: Repeat steps 3 and 4 until a predefined stop condition is satisfied. The stopping condition may be that the number of clusters reaches a predefined value, or the size of the clusters no longer changes significantly, etc.

In the BRICH algorithm, the hierarchical clustering process can be viewed as a bottom-up merging process, where new sub-clusters are formed each time a merger takes place until all sub-clusters are merged to the final root node, forming a complete clustering tree.

In this process, merging between sub-clusters is based on distance or similarity, the smaller the distance or similarity, the more similar the sub-clusters are to each other and the more likely they are to merge together. This bottom-up hierarchical clustering maintains the hierarchical structure of the clusters and does not require a predetermined number of clusters, making it suitable for a variety of types of datasets. Through the process of hierarchical clustering, the BRICH algorithm can get the final clustering results by dividing the data points in the dataset into different clusters and forming a clustering tree with a hierarchical structure.

In BRICH algorithm, for each merge operation, the size of the new cluster after merging is calculated and equalization is controlled according to a predefined threshold. Two sub-clusters Ci and Cj are merged to form a new sub-cluster Ck with size Nk, then the equilibrium control is calculated as follows:

Calculate the size of the new cluster:

$$Nk = Ni + Nj \tag{4}$$

Where Ni denotes the number of data points in subcluster Ci, Nj denotes the number of data points in subcluster Cj, and Nk denotes the data points in the new cluster Ck number.

Threshold setting for equalization control:

Set the predefined upper threshold as Tup and the predefined lower threshold as $Tlow$.

Equilibrium control is performed:

If $Nk > Tup$, it means that the merged new cluster is too large, then continue to split the new cluster Ck into smaller sub-clusters until the size of the new cluster satisfies $Nk \leqslant Tup$ until the size of the new cluster is satisfied. If $Nk < Tlow$, it means that the merged new cluster is too small, then try to merge the new cluster Ck with the neighboring sub-clusters until the size of the new cluster satisfies $Nk \geqslant Tlow$.

Through the above equalization control calculation and adjustment, BRICH algorithm can maintain a relatively balanced cluster size and avoid generating too large or too small clusters, thus improving the quality and stability of clustering. Equalization control is an important feature of BRICH algorithm, which enables the algorithm to better adapt to different data distributions and clustering structures and obtain more reasonable clustering results.

In the BRICH algorithm, stop conditions are used to control the termination of clustering, i.e., when the stop conditions are satisfied, the algorithm does not perform further merging operations but outputs the final clustering results. The stopping conditions include:

The number of clusters reaches a preset value: an expected number of clusters K can be set in advance, and the algorithm stops when the merging operation results in the number of clusters reaching or approaching K.

The size of the clusters no longer changes significantly: a threshold ε can be set such that the algorithm stops when the change in the size of the clusters between two merge operations is less than ε. This ensures that the cluster size is relatively stable.

The similarity of the merge no longer increases: the similarity between clusters before and after the merge operation can be calculated and a similarity threshold can be set. The algorithm stops when the similarity between two merge operations increases less than the threshold.

The error of merging no longer decreases: the error (e.g. average distance, etc.) before and after the merge operation can be calculated and an error threshold is set. The algorithm stops when the error between two merge operations decreases less than the threshold.

We choose to set the size of the cluster no longer changes significantly as a stopping condition for controlling the termination of the BRICH algorithm.

5. Conclusion

During the hierarchical clustering process, record the size of each cluster after each merge operation. The change in the size of the clusters between each merge operation is checked and the difference between the size of the current cluster and the size of the cluster after the last merge operation is calculated. A threshold value ε is set such that when the change in the size of the cluster is less than ε, it is considered that the size of the cluster no longer changes significantly. If the stopping condition is satisfied, the algorithm terminates and outputs the final clustering result; otherwise, it continues with the next merging operation. Finally, we get the clustering results as shown figure 3 below:

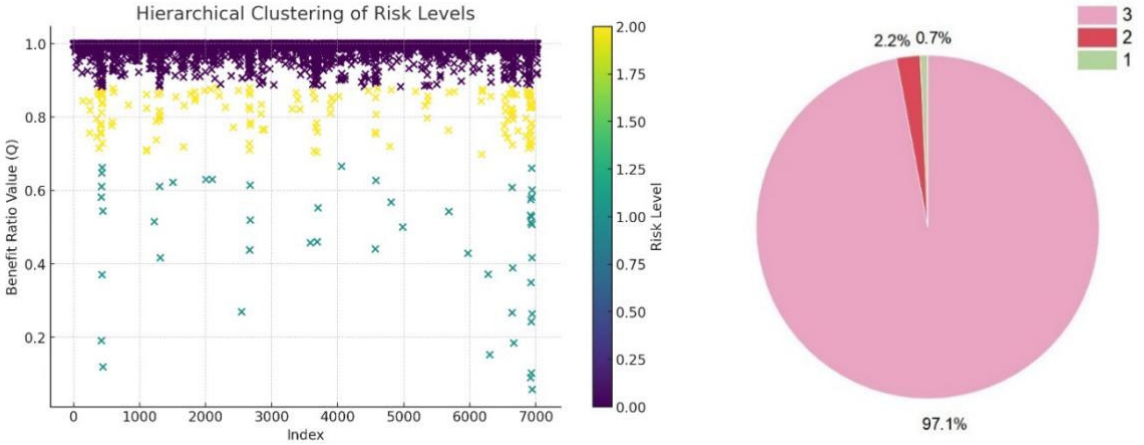

Figure 3. Clustering results

When the Q value is less than 0.7 for clustering category 1, it is a high-risk area, accounting for 0.7%; when the Q value is 0.7-0.9, it is a medium-risk area, accounting for 2.2%; and when the Q value is greater than 0.9, it is a low-risk area, accounting for 97.1%. Different underwriting strategies can be developed for the three different risk areas. Areas with a risk rating of 3: incentives, such as lower premiums, to encourage preventive measures. Areas with a risk rating of 2: Standard underwriting strategies that may include premium adjustments. Areas with a risk rating of 1: may deny coverage or require higher premiums.

We predicted a comprehensive evaluation index of the risk level of the 1976 Tangshan earthquake in China and the 2004 Indian Ocean tsunami data, while categorizing them according to the ratings obtained from clustering, and the final results are shown in the table 4 below.

Table 4. Predicted Risk Levels for the Earthquake in Japan and Hurricane Katrina in the U.S.

Number	Risk indices	Categorization	Number	Risk indices	Categorization
1	0.000559001	Low	4720	0.035966467	Middle
2	0.001253209	Low	4721	0.001586745	Low
3	0.000026065	Low	4722	0.003594854	Middle
4	0.000886683	Low	4723	0.002217961	Low
5	0.000020734	Low	4724	0.014542822	Low
……	……	……	……	……	……
2343	0.001626746	Low	7025	0.086330959	Middle
2344	0.001179073	Low	7026	0.003408208	Middle
2345	0.001254171	Low	7027	0.002058892	Low
2346	0.001231503	Low	7028	0.000035678	Low
2347	0.000228024	Low	7029	0.000000299	Low

References

[1] Cauchois, M., Gupta, S., Ali, A., & Duchi, J. C. (2024). Robust validation: Confident predictions even when distributions shift. Journal of the American Statistical Association, 1-66.

[2] Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., ... & Zheng, B. (2022, October). Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In European conference on computer vision (pp. 612-630). Cham: Springer Nature Switzerland.

[3] Jain, S., Siramshetty, V. B., Alves, V. M., Muratov, E. N., Kleinstreuer, N., Tropsha, A., ... & Zakharov, A. V. (2021). Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods. Journal of chemical information and modeling, 61(2), 653-663.

[4] McCarty, D. A., Kim, H. W., & Lee, H. K. (2020). Evaluation of light gradient boosted machine learning technique in large scale land use and land cover classification. Environments, 7(10), 84.

[5] Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., ... & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature communications, 11(1), 4540.

[6] Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., ... & Xu, H. (2022). Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. Advances in Neural Information Processing Systems, 35, 26418-26431.

[7] Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., & Jiang, S. (2022). Logodet-3k: A large-scale image dataset for logo detection. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(1), 1-19.

[8] Liu, C., Wu, D., Li, Y., & Du, Y. (2021). Large-scale pavement roughness measurements with vehicle crowdsourced data using semi-supervised learning. Transportation Research Part C: Emerging Technologies, 125, 103048.

[9] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine, 4(1), 86.

[10] Chen, L., Lu, Y., Sheng, Q., Ye, Y., Wang, R., & Liu, Y. (2020). Estimating pedestrian volume using Street View images: A large-scale validation test. Computers, Environment and Urban Systems, 81, 101481.