Proceedings of the Sixth International Conference on
Railway Technology: Research, Development and Maintenance
Edited by: J. Pombo
Civil-Comp Conferences, Volume 7, Paper 11.11
Civil-Comp Press, Edinburgh, United Kingdom, 2024
ISSN: 2753-3239, doi: 10.4203/ccc.7.11.11

Freight Train Derailment Severity Estimation using Clustering and Machine Learning Techniques

Z. Saghian[1] and M. Bagheri[2]

[1]School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
[2]School of Railway Engineering, Iran University of Science and Technology, Tehran, Iran

Abstract

This study aims to estimate freight train derailment severity using the U.S. FRA rail accident database spanning from 1997 to 2023. After preprocessing, which included data cleaning and normalization, the dataset comprised 3967 records. The data was split into training (80%) and testing (20%) sets. Using the NBclust function in the R programming environment, optimal clustering for causes was determined, resulting in four clusters based on specific criteria. Each cluster was analyzed using four machine learning techniques: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Gradient Boosting. The results were aggregated based on cluster weights derived from the clustering process. Performance evaluation metrics, including RMSE and Accuracy, were used to assess the models. The findings indicate that all classifiers performed well, with KNN demonstrating superior performance, achieving an accuracy of 92.36% and an RMSE of 0.26. Additionally, the proposed model's average accuracy of 91.53% outperforms the previous benchmark study, which reported an average accuracy of 79.56%. These results suggest that the proposed model is effective in estimating derailment severity and can be a valuable tool for railway safety management.

Keywords: Train Derailment, Clustering, Machine Learning, Freight Train, K-Nearest Neighbors, Support Vector Machines, Random Forest, Gradient Boosting.

1 Introduction

Rail transportation has attracted significant attention from governments and policymakers due to its cost-effectiveness, reliability, high capacity, and favorable environmental impact. However, despite these advantages, rail incidents often result in substantial and irreparable damages. Train derailments, in particular, inflict significant economic losses on countries annually, alongside enforcing pressure on governments to address safety concerns. In both the United States and Canada, train derailments constitute the most common and costly type of rail accident, causing damage to vehicles, tracks, infrastructure, and human lives [1]. Similarly, train derailments remain a serious issue in Europe, with Germany experiencing a high number of rail accidents in 2019 [2] Notably, one of Iran's most devastating derailment accidents in February 2005 led to a tragic loss of lives and injuries [3]. Freight trains involved in rail accidents typically comprise more than 70 cars, resulting in derailments ranging from single locomotives or cars to multiple cars in a single incident [4]. In contrast, only a small percentage of severe highway accidents involve more than three cars [5]. Furthermore, railway accidents present a higher number of potential failure modes due to the size of trains and the complex interactions between equipment and infrastructure. In quantitative safety and risk analyses in the rail transport sector, these factors must be carefully considered to accurately measure the impact of different accident causes, evaluate potential solutions' efficiency, and assess their risk implications.

Train derailment analysis is critical for enhancing rail transportation safety. Understanding the point of derailment (POD) and the number of derailed cars is essential for planning effective safety strategies. Previous studies have investigated factors influencing train derailment and derailment severity. Firstly, Nayak et al. [6] established a statistical relationship between speed and the average number of derailed cars using FRA data, identifying various influencing factors such as accident cause, track type, and train speed. Subsequently, Liu et al. [7] analyzed rail accidents based on causes and predictive parameters, and found that railway tracks accounted for a significant portion of derailment accidents. Furthermore, Liu et al. [8] examined the consequences of derailment by analyzing cause-specific derailments' frequency and severity. Subsequent studies by Liu et al. [9, 10] further investigated derailment rates and severity, identifying covariates affecting derailment rate estimation and developing regression models for severity estimation. The POD, defined as the position of the first derailed car or locomotive in a train, significantly influences derailment probability and severity [11, 12]. Saccomanno and El-Hage [11] developed a geometric model for predicting the number of derailed cars, emphasizing the importance of considering the position of cars within a train. Bagheri et al. [3] introduced a framework for calculating in-transit risk and optimizing train car arrangements to minimize derailment risk. their model is proposed to estimate the probability of derailment by position, using the estimated POD and the number of derailed cars. Liu et al. [9] and Li et al. [13] developed models to estimate derailment severity, highlighting factors such as train speed, residual train length, and train type's

impact. These studies contribute to a comprehensive understanding of train derailment dynamics and inform safety enhancement strategies in rail transportation.

Lotfi et al. [14] also criticized previous studies for their reliance on geometric distributions, which assume independence in each car's derailment and a consistent probability of derailment along the train. They argued that these assumptions do not accurately reflect real-world conditions. Additionally, the failure to consider normalized train length in these studies fails to capture the dynamic forces affecting the train, leading to instability in car-track interactions and train derailment. To address these limitations, Lotfi et al. [14] developed AI techniques to identify causal factors of freight train derailment and model freight train derailment severity. They focused on identifying factors influencing the severity of freight train derailments. They employed various classification methods, including decision trees, random forests, support vector machines, and AdaBoost techniques. Their study revealed that the decision tree emerged as the most effective classifier for predicting derailment severity using the US accident database. Notably, a two-level severity scenario (one car derailed or more) yielded superior classification results. The primary factors affecting derailment severity were identified as train speed, accident cause, and train weight-to-length ratio, with accident cause playing a crucial role in classifying severity.

The primary objective of this study is to develop a new method, following the methodology outlined by Lotfi et al. [14], to predict freight train derailment severity. By incorporating a two-level severity scenario, our aim is to reduce risk and enhance safety in rail transportation while striving for more precise results. To achieve this, we employ four common machine learning approaches along with a clustering method.

2 Methods

In recent years, researchers have increasingly turned to Artificial Intelligence (AI) techniques to investigate train derailments. Compared to traditional methods, AI offers the potential for more precise predictions. By applying data analysis, machine learning, computer vision, and other AI techniques, researchers aim to prevent, detect, and respond to train derailment incidents more effectively. In a study by Huang et al. [15], a novel systematic approach named IG2-LZs was proposed to analyze the causes of passenger train derailment accidents from a management and control perspective. This approach employs a Fault Tree analysis to examine all potential accident causes, with the weight and attribute value of each cause determined using IG2 and LZs, respectively. Bridgelall and Tolliver [16] utilized a clustering approach along with 11 different machine learning models and Principal Component Analysis (PCA). Their findings highlighted strong associations between train derailment and lower track classes, non-signalized territories, and movement authorizations within restricted limits.

As previously discussed, assessing the severity of incidents is a crucial aspect of train derailment analysis. The proposed methodology, outlined in Figure 1, comprises several key steps. Firstly, the data preparation phase involves gathering and cleaning data from the U.S. FRA rail accident database spanning from 1997 to 2023.

Specifically, derailment incidents on main lines involving trains with a maximum of three head-end locomotives are considered. Ten relevant features which may influence the severity of a derailment are selected, including train speed, gross tonnage of freight, point of derailment (POD), number of head-end locomotives, derailment cause, temperature, visibility, weather, and track class. There are five cause group categories for the 389 distinctive causes listed in the FRA database, which include track, roadbed, and structure (T); signal and communication (S); mechanical and electrical failures (E); train operation-human factors (H); and miscellaneous (M) [3]. Causes are identified by codes beginning with these letters followed by three digits, such as T205 or S003. For this study, human factor and miscellaneous causes were excluded.

After cleaning and preprocessing the data, suitable features are identified for input into the models. This study aims to introduce a novel method for estimating the severity of freight train derailments. To achieve this goal, a clustering technique is applied in conjunction with support vector machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbours (KNN) algorithms, which are recognized as powerful machine learning approaches.

The remaining data is then subjected to a clustering method, which comprises a combination of various clustering algorithms. The POD is clustered based on different causes, and the data is subsequently partitioned according to the obtained clusters. Next, each machine learning algorithm is employed to model the severity of derailment within each cluster.

Consistent with Lotfi et al. [14], a two-level severity scenario, distinguishing between incidents involving one car derailed or more, is applied to each cluster. The two-level severity of derailment is determined for each cluster, and the final result is computed by considering the prediction results of different clusters, weighted accordingly. Subsequently, the final result for each machine learning algorithm is obtained. These results are then compared both amongst themselves and with the findings of Lotfi et al. [14]. Overall, this methodology enables a comprehensive assessment of train derailment severity, considering both clustering techniques and machine learning algorithms to achieve more accurate and insightful results.

The machine learning algorithms are briefly described.

2.1 Support Vector Machine (SVM)

Support Vector Machines (SVM) utilize the structural risk minimization principle to achieve robust generalization with limited training data. Originally developed by Vapnik [17] and colleagues at AT&T Bell Laboratories, SVMs are proficient in discerning intricate patterns within complex datasets. This learning algorithm excels in discriminative classification by extrapolating from known examples to predict classifications for unseen data [18, 19].

2.2 Decision Tree

A decision tree embodies a hierarchical, flowchart-like structure in which each internal node denotes a test on a particular attribute. The outcomes of these tests are represented by branches, leading to leaf nodes that signify class labels. When presented with a data instance represented as a tuple X, the decision tree assesses the attribute values of X against its structure, following a path from the root to a leaf node.

At the leaf node, the decision tree assigns a class prediction for the given tuple. Decision trees are highly interpretable and can easily be transformed into classification rules. In the field of predictive modelling, decision trees are particularly adept at handling classification tasks where the target variable assumes a finite set of values. Notably, decision trees can be constructed more quickly than other classification methods [20].

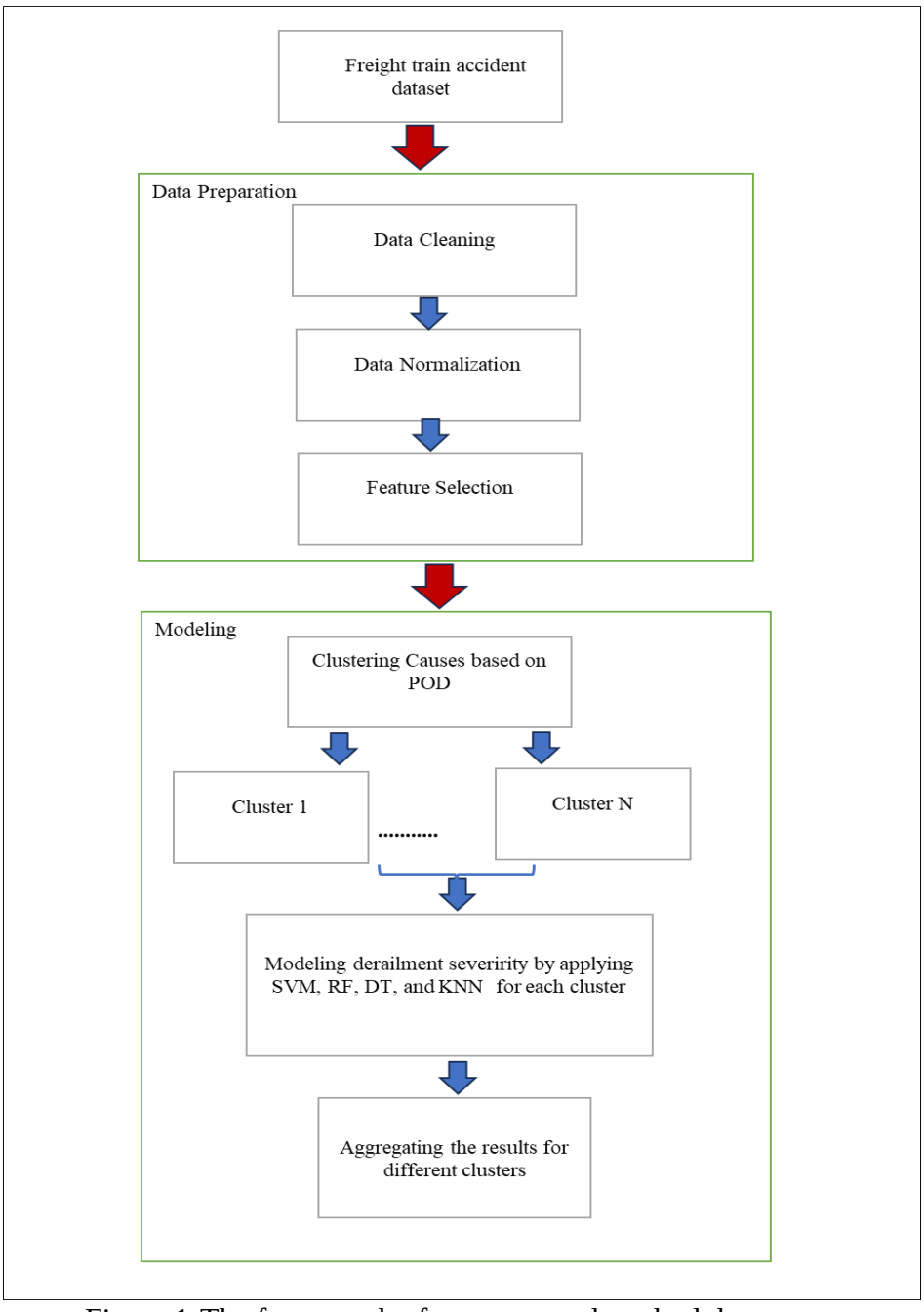

Figure 1-The framework of our proposed methodology

2.3 Random Forest

Random Forest, a widely used machine learning technique for developing prediction models. Random Forest, initially introduced by Breiman in 2001 [21], comprises a collection of classification and regression trees. While decision trees offer an intuitive method for prediction, they often lack accuracy for complex datasets. Random Forest overcomes this limitation by constructing multiple trees using randomly selected training datasets and predictor variables, and aggregating results for higher prediction accuracy. Notably, Random Forest maintains the interpretability of decision trees while offering superior performance. It is recognized for its ability to handle large datasets with numerous predictor variables efficiently. Random Forest consistently achieves high prediction accuracy compared to other models, making it a preferred choice in classification settings. Each decision tree within the Random Forest predicts the class of input, and the final prediction is determined by the class receiving the most "votes" from all trees [22].

2.4 K nearest neighbour (KNN)

The k Nearest Neighbours algorithm (KNN) is an instance-based or lazy learning method, widely recognized as one of the simplest machine learning algorithms. Its premise lies in the assumption that similar samples belonging to the same class are highly probable. The core concept of the KNN algorithm entails selecting the k nearest neighbours for each test sample and subsequently using these neighbours to predict the class of the test sample. Due to this approach, the KNN algorithm is often considered to require no explicit training step. Nonetheless, it remains a popular classification method in data mining and statistics, owing to its straightforward implementation and noteworthy classification performance [23].

2.5 Performance Evaluation

The dataset is divided into training and test sets. The parameters of the model are tuned using the training dataset. To estimate the performance of our proposed method, we adopt the root mean square error (RMSE). We also take into account the accuracy measure. Then we compare the results of our proposed method with the results of a Lotfi et al. [16] study as a benchmark.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where y_i and $\hat{y}_i$ are the output observations, and the model prediction, respectively. The number of observational samples equals *N*.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

The examination is performed within the R programming environment, where the implementation of the codes also took place. For the clustering phase, the NBclust function in R was utilized. This function aids in determining the optimal number of clusters by evaluating 30 distinct indices such as silhouette width, Davies-Bouldin index, and gap statistic. Through the variation of cluster numbers, distance metrics,

and clustering techniques, Nbclust explores diverse configurations to identify the most suitable solution.

3 Results

In this study, the FRA database is utilized to assess freight train derailment severity. Following a data preprocessing procedure, certain rows and columns are excluded. Ultimately, the dataset comprises 3967 records. The training dataset encompasses 80% of this data, with the remainder allocated to the test set. Initially, the NBclust function within the R programming environment is employed to identify the optimal clusters for the causes. The clustering outcomes are depicted in Figure 2.

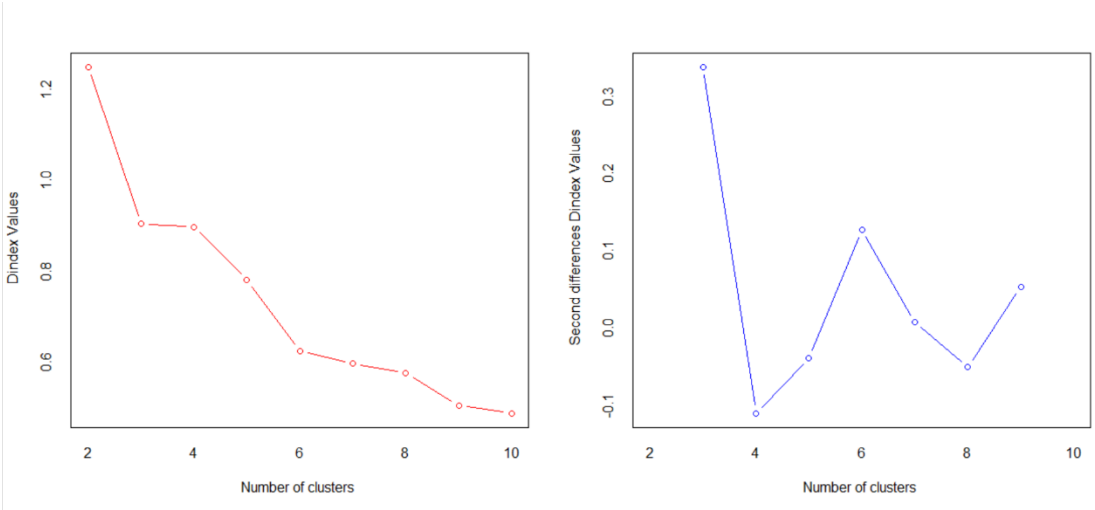

Figure 2- applying NBclust function to discover efficient number of clusters

The analysis shows the optimal number of clusters to be four. Subsequently, each of the four machine learning techniques is applied to each cluster. The results for each technique across different clusters are aggregated, taking into account their respective weights (calculated as the ratio of each cluster's population to the total number of data points). Performance evaluation is conducted using RMSE and Accuracy measures, with the results summarized in Table 1.

Table 1: The performance of our proposed methodology

Classifier	RMSE	ACCURACY
SVM	0.260569	0.921281
RF	0.266954	0.913859
DT	0.268983	0.910875
KNN	**0.260465**	**0.923602**

As illustrated in Table 1, all four classifiers demonstrate excellent performance, with KNN exhibiting superior results compared to the others. Subsequently, the performance of our proposed method is compared to that of the study by Lotfi et al. [16] in terms of Accuracy, with the results presented in Table 2.

Table 2 Comparing the results in term of Accuracy

Classifier	Lotfi et al. (2022)	Our proposed method
SVM	0.801	**0.921281**
RF	0.783	**0.913859**
DT	0.803	**0.910875**

As shown in Table 2, the results of our proposed model surpass the benchmark in terms of Accuracy measure.

4 Conclusions and Contributions

In conclusion, this study successfully developed a methodology to predict freight train derailment severity, utilizing machine learning techniques and data from the FRA database. Through complicated preprocessing and analysis, optimal clusters for causes were identified, facilitating accurate severity estimation. The results highlight the efficacy of machine learning models, particularly KNN, in predicting derailment severity. Furthermore, our proposed model demonstrated superior accuracy compared to the benchmark study. Future research could explore additional factors and refine the methodology for even more precise predictions.

Acknowledgements

This work is based upon research funded by Iran National Science Foundation and Iranian Supreme Council for Science Research & Technology under project No.4022798 (INSF)

References

[1] Federal Railroad Administration (2018). Railroad accident database. https://railroads.dot.gov/safety-data/accident-and-incident-reporting/train-accident-reports/train-accident-reports

[2] Report on Railway Safety and Interoperability in the EU. 2020. https://www.saferail.nl/ERA/ERAPUBLICATIONS/2020_Report-Railway-Safety+Interoperability.pdf

[3] M. Bagheri, F. Saccomanno, S. Chenouri, L. Fu. (2011). Reducing the threat of in-transit derailments involving dangerous goods through effective placement along the train consist, Accident Analysis & Prevention. 43(3) 613-620.

https://doi.org/10.1016/j.aap.2010.09.008
[4] AAR (Association of American Railroads). (2018). “Railroad facts.” Accessed December 21, 2018. https://my.aar.org/Pages/Product-Details .aspx?ProductCode=RFB2018Web.
[5] NHTSA (National Highway Traffic Safety Administration). (2008). National motor vehicle crash causation survey: Report to Congress. Washington, DC: NHTSA.
[6] P.R. Nayak, D.B. Rosenfield, & J.H. Hagopian. (1983). Event Probabilities and Impact Zones for Hazardous Materials Accidents on Railroads. US Department of Transportation, Federal Railroad Administration DOT/FRA/ORD-83/20. Washington, DC, USA.
[7] Q. Liu, H. Liu, X. Xie & Y. Li. (2011). Analysis on derailment accident causes and prediction. Procedia Engineering, 15, 1390–1394.
[8] X. Liu, M. R. Saat & C. P. L. Barkan (2012). Analysis of Major Derailment Causes on Heavy Haul Railways in the United States.
[9] X. Liu (2013). Optimal Strategies to Improve Railroad Train Safety and Reduce Hazardous Materials Transportation Risk. Doctoral Thesis, University of Illinois at Urbana Champaign, Department of Civil and Environmental Engineering. Urbana, IL, USA.
[10] X. Liu, M. Rapik Saat & C. P. L. Barkan. (2017). Freight-train derailment rates for railroad safety and risk analysis. Accident Analysis & Prevention, 98, 1-9. https://doi.org/https://doi.org/10.1016/j.aap.2016.09.012
[11] F.F. Saccomanno, & S.M. El-Hage (1989). Minimizing derailments of railcars carrying dangerous commodities through effective marshaling strategies. Transportation Research Record: Journal of the Transportation Research Board, 1245: 34-51.
[12] F.F. Saccomanno, & S.M. El-Hage. (1991). Establishing derailment profile by position for corridor shipments of dangerous goods. Canadian Journal of Civil Engineering, 18(1): 67-75.
[13] X. Liu, M. Rapik Saat, & C. P. L. Barkan. (2024). Statistical Analysis of Train Derailment Severity for Unit Trains Versus Manifest Trains. Transportation Research Record, 2678(4), 30-42. https://doi.org/10.1177/03611981231182989
[14] A. Lotfi, M. Bagheri, & A. Ahmadi. (2023). Using Machine Learning Methods for Modeling Freight Train Derailment Severity. Transportation Research Record 2677(3): 961–73.
[15] W. Huang, Y. Zhang, D. Yin, B. Zuo, M. Xu, & R. Zhang, (2021). Using improved Group2 and Linguistic Z-numbers combined approach to analyze the causes of railway passenger train derailment accident. Information Sciences, 576, 694–707. https://doi.org/10.1016/j.ins.2021.07.067
[16] R. Bridgelall, and D. Tolliver (2021). Railroad Accident Analysis Using Extreme Gradient Boosting. Accident Analysis and Prevention, 156(2021). DOI: 10.1016/j.aap.2021.106126, 2021(106126).

[17] V. Vapnik, S.E. Golowich, & A.J. Smola. (1997). Support vector method for function approximation, regression estimation, and signal processing. In Advances in Neural Information Processing Systems (pp. 281-287).
[18] C. Cortes, & V. Vapnik. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
[19] D. Basak, S. Pal, D. Ch, & R. Patranabis. (2007). Support Vector Regression. Neural Information Processing—Letters and Reviews, 11, 203-224.
[20] L. Breiman, J.H. Friedman, R.A. Olshen, & C.J. Stone. (1984). Classification and Regression Trees. Chapman and Hall/CRC.
[21] L. Breiman. (2001). Random Forests. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324
[22] A. Palczewska, J. Palczewski, R. Marchese Robinson, & D. Neagu. (2014). Interpreting Random Forest Classification Models Using a Feature Contribution Method. In: Bouabana-Tebibel, T., Rubin, S. (eds) Integration of Reusable Systems. Advances in Intelligent Systems and Computing, vol 263. Springer, Cham. https://doi.org/10.1007/978-3-319-04717-1_9
[23] D. Cheng, S. Zhang, Z. Deng, Y. Zhu, & M. Zong. (2014). KNN Algorithm with Data-Driven k Value. In: Luo, X., Yu, J.X., Li, Z. (eds) Advanced Data Mining and Applications. ADMA 2014. Lecture Notes in Computer Science, 8933. Springer, Cham. https://doi.org/10.1007/978-3-319-14717-8_39