

# AUTOMATED CLASSIFICATION AND CODING FOR BIM COMPONENTS BASED ON APPLICABLE BIG DATA

Xinglei Xiang<sup>1</sup>, Zhiliang Ma<sup>1</sup>, Jiayi Li<sup>1</sup>

*1 Department of Civil Engineering, Tsinghua University, Beijing 100084, China*

## Abstract

For the application of BIM technology, BIM models play a vital role in transferring and sharing building information. BIM models represent the building information through the attributes of components, such as walls and slabs, and the relationship between these components. Among all the attributes of components in BIM models, the classification and coding attribute is essential to retrieve the building information in a quick way. However, in practice, the BIM models that are prepared by using a BIM authoring tool cannot ensure complete and correct classification and coding attribute when they are transferred into the format of another tool. Besides, BIM models prepared by using 3D reconstruction technology also lack the classification and coding attribute. Missing or incorrect classification and coding attribute of BIM components impedes the fully exploitation of BIM model greatly. To solve the problem, this paper proposes an automated classification and coding method for BIM components based on a batch of BIM models with components labeled with key features and type, which can be obtained from the big data of the BIM models that have been correctly used. In the method, the association rule mining algorithm is used to establish the classification and coding rules for BIM components based on the labeled component data set. Then, for any BIM component, an algorithm based on the credibility reasoning approach is used to execute the rules and obtain its classification and coding attribute. In this way, the classification and coding attribute can be determined for any BIM component according to any given standard. The method is validated by developing a prototype based on Autodesk Revit and by using a batch of BIM models from structural design, and 92.7% precision is achieved in the test case. This method contributes to the quick classification of BIM components.

© 2023 The Authors. Published by Diamond Congress Ltd.

Peer-review under responsibility of the scientific committee of the Creative Construction Conference 2023.

**Keywords:** Building information modeling, association rule mining algorithm, classification and coding.

## 1. Introduction

BIM (Building Information Modeling) is an essential carrier of information for the entire lifecycle of a construction project. The BIM model database is the core resource for BIM applications in construction projects [1]. Standardized classification and coding of the components in BIM models are prerequisites for accurately retrieving and accessing components in BIM models. This is also a necessary condition for BIM technology to be used for supporting the sharing of building information data throughout the entire lifecycle of a construction project [2]. For example, because various construction robots have

different construction targets, a wall polishing robot only needs wall data from the building information model. Therefore, both the accurate retrieval and access of wall model components from the structural BIM model, and the transmitting of the wall data to the wall polishing robot are prerequisites for the smooth implementation of subsequent work.

To standardize the classification and coding of BIM components and facilitate the exchange and sharing of information throughout the entire lifecycle of a construction project, some countries have established building classification and coding standards, such as Unifomat II, Omni Class and so on. To promote the rapid development of China's construction industry, a series of building product classification and coding standards have also been established, including "Standard for classification and coding of building information model (GB/T 51269-2017)" and "Classifying and coding of construction products (JG/T 151-2015)". Among them, the former is a standard related to the IFD (International Framework for Dictionaries) and was established by referring to OmniClass. It is applicable to the classification and coding of BIM models in civil buildings and general industrial buildings. This standard, which is intended for the construction engineering field, stipulates the standard for classification and coding for various types of information. The coding structure includes table codes, major category codes, intermediate category codes, minor category codes, and detailed category codes, with each level of code represented by two Arabic numerals. For example<sup>[3]</sup>, the code 14-10.20.36.06 represents the code for outdoor stairs, with the table code being 14, the major category code being 10, the intermediate category code being 20, the minor category code being 36, and the detailed category code being 06.

As a carrier of building information, BIM models are transferred and utilized among different modelling application software, including building energy simulation software and cost estimation software, which requires accurate classification of components in a BIM model. Failure to properly classify the BIM components during the modelling process, loss of information during storage and transmission, or inability to recognize during component information extraction can result in missing classification information for BIM components, which in turn impede the subsequent model information processing work<sup>[4]</sup>. In addition, with the help of BIM software, it is relatively easy to perform component classification and add semantic information during the modelling process<sup>[5]</sup>. Nevertheless, a 3D model created from point cloud data contains only geometric information and lacks semantic information<sup>[6]</sup>. Therefore, the problem of missing or incorrect classification and coding attribute of components in the obtained BIM models is widespread and significant, which severely impedes the further application of BIM technology in various tasks and stages, such as design, construction and operation. Moreover, due to the complexity of the classification and coding standards for components in a BIM model, the variety and large number of designed components, the cost of manual classification and coding is high and prone to errors. In response to these problems, it is obvious that automatic classification and coding tools are needed.

Consequently, this study aims to establish an automatic classification and coding method for structural components in BIM models, and develop a plugin based on an authoring tool of BIM model, taking

Autodesk Revit as an example, for customization to add classification and coding attributes to components in structural BIM models automatically, in which the classification and coding attribute of these components are obtained through rule reasoning. The rules used for the reasoning are obtained from a labelled data set through data mining algorithms. Through the method proposed in this study, the classification and coding attributes of structural components in BIM models can be replenished automatically, which contributes to promote the deepening application of BIM.

## 2. Related works

There are mainly two types of researches related to automatic classification and coding methods for components in a BIM model, i.e., those based on machine learning and those based on rule-based reasoning.

In the methods based on the machine learning, the overall process can be summarized as follows [7,8]. Firstly, a portion of components in a BIM model in the data set are selected, and their semantic information, such as name, code, type, size, and fill ratio, is manually labelled as the training set for the machine learning algorithm. After that, based on the selected data set, the machine learning model is trained with the labelled data set to predict the coding information of unlabelled objects, by using methods such as decision trees, neural networks, and support vector machines. After training, the model needs to be tested to see if it can perform well on the test set. Ultimately, once the model is evaluated and deemed satisfactory, it can be used to fill in the semantic information of unlabelled components automatically and thus add the classification and coding attribute for components in a BIM model automatically [9]. According to the above process, Bloch et al. [10] trained an automatic room classification model using a multiclass neural network algorithm on the AZURE ML platform, with 150 BIM model instances as samples, 70% of which is used as the training set, and 30% is used as the testing set. A multiclass neural network algorithm is used for spatial classification of residential apartments. The results indicated that the machine learning-based method was useful and efficient for room classification. Koo et al. analysed the geometric dimensions and spatial relationships of these components, based on a data set of 4187 independent components in a BIM model from six architectural BIM models. Next, identified unconventional or outlier data that may require further analysis to improve the reliability of the component classification model, through using support vector machine and introducing a novelty detection algorithm. The evaluation results indicated the effectiveness of the algorithm for automatic classification and coding of components in a BIM model [11]. As described above, the structural components in a BIM model can be classified and encoded through machine learning methods. Objects such as tables, chairs, safety cones, etc. in construction sites, office spaces, and other settings can also be identified and classified using machine learning methods. Ferguson et al. designed an algorithm that can be used for both 2D image data and 3D point cloud data to recognize objects such as trash cans and safety cones at construction sites. The study collected 1214 RGB-D image data as the training data set and 255 RGB-D image data as the testing data set to conduct object classification experiments, which was verified to be available for classification and coding of components [12]. As

mentioned above, different algorithms are required to classify and code components in a BIM model in different scenarios based on machine learning method. Meanwhile, it is necessary to manually establish large-scale data set and label them, which is highly dependent of manual operations.

The rule-based inference methods are widely used for the classification and coding of components in a BIM model. Firstly, formulating rules are found and selected, through the analysis of the relevant knowledge of components in a BIM model. Each component in the BIM model would be classified, after being identified according to the rules. For the components that cannot be identified based on rules, their information can be manually marked to supply the rules; finally, the classification and coding results would be validated and corrected to ensure their accuracy and consistency<sup>[13]</sup> For example, the name of components in BIM models would be checked to ensure accuracy, the size and material would be checked to match the design document, etc. Some researchers established rule sets to infer the classification and coding of BIM models by acquiring the knowledge of domain expert, and adding inferred classification and coding attribute to the components<sup>[14]</sup>. To classify components in a BIM model through a customized matching algorithm, and to provide mathematical measures of matching results at the same time<sup>[15]</sup>. The rule-based method has also been applied to establish rule sets for inferring the classification and coding of components in a BIM model based on their geometric attributes and geometric theorems. This method involves algorithm construction, extraction of objects of components in BIM models, object classification, semantic filling, classification result verification, and algorithm evaluation<sup>[16]</sup>. As illustrated above, the rule-based reasoning method relies on the knowledge of domain experts, and its effectiveness is also influenced by the methods and accuracy of rule content transformation.

To address the limitations of previous studies, this research proposes the data mining approach to classify and code structural components in BIM models as an example. The method aims to mine the classification and coding rules from a large data set of components of BIM models, avoiding the high requirement for domain experts' knowledge and the impact of rule transformation and accuracy on the effectiveness of the method. Firstly, a data set for rule mining is formed by collecting and processing structural BIM models. Then, to apply the association rule mining algorithm to obtain rules for classification and coding automatically. Finally, to introduce the credibility reasoning approach to fully consider the uncertainty of rules, and thus achieve the automatic classification and coding of these components. This method is also expected to be applicable to the automatic classification and coding of other types of components such as MEP components. To achieve the objective the credibility reasoning approach is introduced in this phase.

### **3. Methodology**

The classification and coding method for structural components in BIM models proposed in this paper consists of three phases: data preparation, rule mining, and credibility reasoning. As shown in Fig. 1, this section will provide a detailed explanation of each phase.

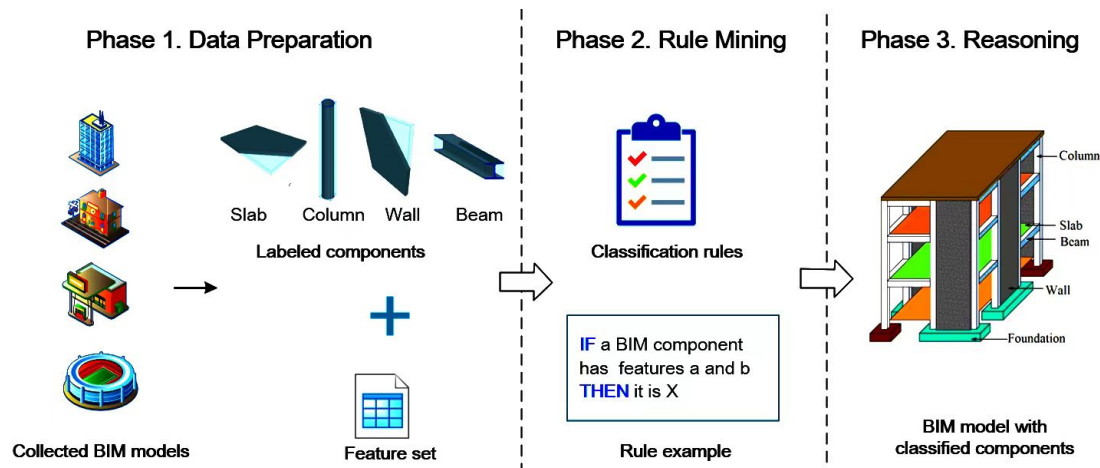


Fig. 1. Process overview of method

(1) Data preparation. The main objective of the phase is to provide the data of structural components in BIM models for rule mining, which includes BIM models collection, and components labelling with beams, slabs, columns, walls, and component features design. Then, the feature design of the component is aimed at obtaining the values of various features of the component, thereby forming a data set for rule mining. The feature design and data set establishment of the component is the prerequisites for rule mining.

(2) Rule mining. The main purpose of this phase is to apply association rule algorithm for rule mining, and to establish a rule set for automatic classification and coding of components in BIM models, including three steps: obtaining frequent item sets from the data set, generating rules from frequent item sets, and controlling the size and quality of the rule sets. The quality and quantity control of the rule sets is the foundation for ensuring the accuracy of classification and coding.

(3) Reasoning. The main objective of this phase is to reduce or eliminate the uncertainty inherent in the rules established by using data-driven methods, including two steps: calculating the internal certainty factor of the rules and calculating the certainty factor of the classification and coding results when a given component feature matches the rules.

#### 4. Data preparation

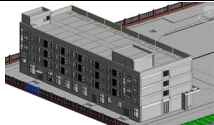
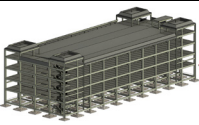
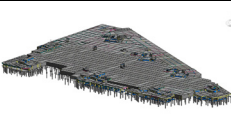
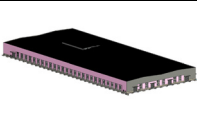
The proposed data mining approach consists of three phases, namely feature design, data set establishment, and rule mining. Considering the requirement on the consistency in the classification and coding standards, this research will be carried out with reference to the “Standard for classification and coding of building information model (GB/T 51269 – 2017)”. The method is expected to be also applicable to other standards.

##### 4.1 Data collection

The data for structural components in BIM models used in this study mainly includes the structural BIM models of four actual building projects. Through the analysis of 8838 components in the BIM models of

these projects, a data set for rule mining was established. The quantity of various types of components for each project is presented in Table 1.

Table 1. Overview of collected BIM models

NO.	Model 1	Model 2	Model 3	Model 4
Image				
No. of beam	311	422	1221	396
No. of column	156	342	1838	375
No. of slab	265	89	1738	22
No. of wall	670	0	896	97

#### 4.2 Feature extraction

The values of various features of structural components in BIM models were obtained, through the development of plugins using the Revit API, thus forming a data set for rule mining. The format of this data set used for rule mining is stored in a standard text file format, which follows the convention of “component category (label) + feature name: feature value”. The features of the components can be mainly divided into two categories: attribute features and relationship features. Attribute features include surface area and volume, while relationship features include component intersection and parallelism. Among them, the attribute feature is usually divided into geometric features and non-geometric features, for example, “aspect ratio” is a geometric feature, and “fill rate” is a non-geometric feature. The value types of features are divided into continuous types and discrete types. For example, the “volume of these components” feature has a continuous value type, while the “whether the component intersects with a column” feature has a discrete value type.

The process of features designing of structural components in BIM models must ensure the effectiveness of the selected component features and their values. For example, the feature “structure material is concrete” is not an effective feature for distinguishing between beams and columns. The feature “longest vertical edge” is an effective feature for distinguishing between columns or walls but not plates. After conducting testing on the geometric dimensions, volume, surface area, material, and other attributes of components in BIM models, it was found that the effectiveness of material, location, and other attribute features in component classification and coding is insufficient and they are not included in the feature set for the classification and coding of components in BIM models in this study. After an extensive literature review and repeated experimental testing, a set of features for the classification and coding of components in BIM models was identified. The geometric features of the properties include aspect ratio, short-to-medium ratio, fill rate, whether it is a hexahedron, surface area, volume, and vertical extension. The relational attributes include whether it intersects with walls, columns, beams, foundations, and floor slabs. The types, names, and meanings of features of these components in a BIM model are presented in Table 2 for reference.

Table 2. Feature definitions

Feature type	Feature name	Explanation
Geometric feature	Short-long ratio	The ratio of the shortest edge to the longest edge of the bounding box of the component.
	Short-middle ratio	The ratio of the shortest edge to the second-longest edge of the bounding box of the component.
	Fill ratio	The fill ratio of the volume of the component to the volume of the bounding box of it.
	Face number	The face number of the component.
	Surface area	Surface area of the component.
	Volume	Volume of the component.
	Vertically extended	Whether the member extends vertically.
Relational feature	Relation with wall	Whether the component intersects a wall.
	Relation with column	Whether the member intersects a column.
	Relation with beam	Whether the member intersects a beam.
	Relation with foundation	Whether the member intersects a foundation.
	Relation with slab	Whether the member intersects a slab.

## 5. Rule mining

Due to the low requirement for domain knowledge for data-driven rule mining methods, the methods are capable of continuously optimizing rules from expanded data. Additionally, rule mining is supported by mature algorithms with strong generality. In this paper, the data mining approach was applied to establish rules for the classification and coding of structural components in BIM models. Based on the obtained data set, the association rule mining algorithm was applied to obtain rules for classification and coding automatically. As a widely used data mining algorithm, the association rule mining algorithm discovers frequent combinations of feature values that co-occur with specific component types in the data set, i.e., frequent item sets, forming “IF-THEN” decision rules. The “IF” part is the antecedent, which expresses a set of feature values of component, while the “THEN” part is the consequent, which expresses the classification and coding result for the component.

The process of rule mining using the association rule mining algorithm can be divided into three steps. Namely, firstly, to extract frequent item sets from the data set, which are statistical patterns of co-occurrence of feature values and component categories with high frequency. Then, to generate rules from frequent item sets, where the condition part of the rules consists of the feature values in the frequent item sets and the result part of the rules corresponds to the component types in the frequent item sets. Finally, to filter the rules to control the size and quality of the rule base and improve the efficiency and effectiveness of subsequent automatic reasoning processes. The selection criteria are divided into two categories: absolute standards and relative standards. Among them, the absolute criteria for rule selection include support and confidence, where support is the probability of frequent item sets appearing in the data set, which ensures the universality of the rule, while confidence is the conditional

probability that the result holds when the condition occurs in the data set, ensuring the quality of the rule. Relative standards for rule selection involve comparisons within the rules themselves. For example, when comparing Rule A and Rule B with the same result, if the constraint conditions decrease while the confidence of the result increases, it indicates that compared to Rule A, Rule B introduces an additional constraint, which lowers the probability of the result being true. In that case, Rule A is better than Rule B, and Rule B should be eliminated from the rule sets to avoid conflicts within the rules. Here is an example,

(1) Rule A: If the longest edge is vertical (FALSE), the ratio of short to medium edges is greater than 0.4 and not less than 0.6 (TRUE), the connected base number is 0 (TRUE), then the component is judged to be a reinforced concrete beam with a confidence level of 0.994.

(2) Rule B: If the longest edge is vertical (FALSE), the ratio of short to medium edges is greater than 0.4 and not less than 0.6 (TRUE), the connected base number is 0 (TRUE), the surface area is less than 10 square meters (TRUE), then the component is judged to be a reinforced concrete beam with a confidence level of 0.954.

Compared with Rule B, Rule A has fewer constraints but a higher confidence level for the same result. Therefore, Rule A is better than Rule B. In Rule B, an additional constraint on the surface area is introduced, which lowers the confidence level of the rule, and therefore, Rule B will not be included in the rule sets for the reinforced concrete beam.

According to the test results, the threshold for support is set as 0.05 and that confidence as 0.8. Frequent item sets are obtained from the data set, and then rules are generated from the frequent item sets, and the rules are filtered finally. Among them, the filtering requirements are as follows: support > 5%, confidence > 80%, and there is no better rule. A total of 114 rules for columns, walls, beams, slabs, and other structural components in BIM models are established. Taking the rules for reinforced concrete beams as an example, the classification and coding rules for beams obtained by using this algorithm are shown in Table 3.

Table 3. Classification and coding rules for reinforced concrete beams established using the data mining approach

No.	Rule
1	IF the vertical edge is the longest (FALSE) AND the ratio of short to medium sides is greater than 0.4 and not less than 0.6 (TRUE) AND the connected bases number is 0; THEN it is a reinforced concrete beam.
2	IF the vertical edge is the longest (FALSE) AND the ratio of short to medium edges is greater than 0.4 and not less than 0.6 (TRUE); THEN the component is a concrete beam.
3	IF the vertical edge is the longest (FALSE) AND surface area is less than 10 square meters (TRUE) AND number of connected columns is 0 (TRUE) AND the connected base number is 0 (TRUE); THEN it is a reinforced concrete beam.
4	IF the vertical edge is the longest (FALSE) AND surface area is less than 10 square meters (TRUE) AND number of connected bases is 0 (TRUE); THEN it is a reinforced concrete beam.
5	IF the vertical edge is the longest (FALSE) AND surface area is less than 10 square meters (TRUE) AND number of connected columns is 0 (TRUE); THEN it is a reinforced concrete beam.
6	IF the vertical edge is the longest (FALSE) AND surface area is less than 10 square meters (TRUE); THEN it is a reinforced concrete beam.

7 IF the ratio of short side to medium side is greater than 0.4 and not less than 0.6 (TRUE) AND number of connected bases is 0 (TRUE); THEN it is a reinforced concrete beam.

## 6. Reasoning

According the pre-established rule base, a component of unknown type is classified by matching its features with the condition clauses of these rules. Considering the uncertainty of rules established by using data-driven methods, this study introduces the credibility reasoning approach to ensure the completeness, consistency, and rationality of the classification process.

In the theory of the credibility reasoning approach, the uncertainty of rules in the form of “IF E, THEN H” is evaluated by the parameters of certainty factor (CF), including the three types, i.e., CF(H, E), CF(E) and CF(H). CF(H, E) is used to evaluate the uncertainty of the logical relationship between the conditions (E) and result (H). Besides, the uncertainty of the conditions in a rule also needs to be considered. For example, in most cases, the qualitative conditions have higher uncertainty than quantitative conditions. To handle the uncertainty, CF(E) is a parameter to evaluate the matching degree between the given component feature values and the conditions with uncertainty. Apparently, for a condition without uncertainty, the value of CF(E) is TRUE (1) or FALSE (0), in which 0 means not matching, and 1 means matching. CF(H) indicates the reliability when a result is deduced. Table 4 gives the definitions and explanations of these parameters.

In the classification and coding method based on the credibility reasoning, CF(H) is the decisive indicator to draw the result. Namely, after matching the features of a structural component with all the rules, the final type of the component is the result of the matched rule with the highest CF(H). Thus, the calculation processes of CF(H) are introduced. Firstly, the base case is considered when only one rule is matched. On this basis, the calculation process when multiple rules are matched is given <sup>[17]</sup>.

Table 4. Definitions and explanations of parameters

Parameter	Definitions	Explanations
E	The conditional part of the classification and coding rules.	Production rule: IF E, THEN H.
H	The result part of the classification and coding rules.	Production rule: IF E, THEN H.
CF(E)	The certainty factor of condition E.	The degree of matching between the given component feature values and the conditions E in the rules. The value is either 0 or 1, where 0 means rule matching failure, and 1 means rule matching success.
CF(H, E)	The certainty factor of the classification and coding rule “IF E, THEN H”.	To reflect the strength of the relationship between the premise condition E and the result H.
CF(H)	The certainty factor that the result H holds under the condition of matching the component characteristics with the rule.	The final rule-based inference result is the one with the highest degree of certainty factor.
P(H)	The probability of result H being true.	To be calculated by statistical analysis of 8,838 components in the data set.
P(H E)	The conditional probability of the result of H being true when the condition E is true, namely, the confidence.	To be calculated from the results of rule mining.

If only one rule is matched, according to theory of the credibility reasoning approach, the  $CF(H)$  depends on the value of  $CF(H, E)$  and  $CF(E)$ , as shown in Equation (1).

$$CF(H) = CF(H, E) * CF(E) \quad (1)$$

Because the conditions of the classification and coding rules are quantitative and has no uncertainty, the value of  $CF(E)$  is TRUE (1) or FALSE (0) and can be obtained directly from the matching result. Thus, the value of  $CF(H)$  can be obtained by calculating the value of  $CF(H, E)$  by using Equation (2).

$$CF(H, E) = \begin{cases} \frac{P(H|E) - P(H)}{1 - P(H)}, & \text{if } P(H) \neq 1 \\ 1, & \text{if } P(H) = 1 \end{cases} \quad (2)$$

Where  $P(H|E)$  is the probability of result  $H$  when the condition  $E$  is matched and it is equal to the confidence of a rule in the rule mining process according to the definition.  $P(H)$  is the probability for “the result is true”. Specific to the classification and coding rules, the result of a certain component type and the  $P(H)$  can be obtained by counting the frequency of component types in the data set.

Based on the calculation equation of  $CF(H)$  in the case that only one rule is matched, when  $n$  ( $n > 1$ ) rules with the conclusion  $H$  are matched, the  $CF(H)$  can be calculated by using an iterative process, as shown in Fig. 2. As the known conditions of the process,  $n$  rules with the conclusion  $H$  are matched and the certainty factors of the  $i$ -th ( $1 \leq i \leq n-1$ ) rule of the  $n$  rules, i.e.,  $CF_i(H)$ , can be obtained by the Equation (1). In the first step of the process,  $CF(H)$  is initialized by  $CF_1(H)$ . Then, in each iteration of the process, the other rules are handled one by one and the value of  $CF(H)$  is updated by using Equation (3).

$$CF(H) = CF(H) + CF_i(H) - CF(H) * CF_i(H) \quad (3)$$

Thus, after  $n-1$  iterations, the final  $CF(H)$  is calculated and the iterative process ends<sup>[15]</sup>. Note that in the process, the calculation result of  $CF(H)$  is independent of the sequence of the matched rules due to the commutative laws of Equation (3).

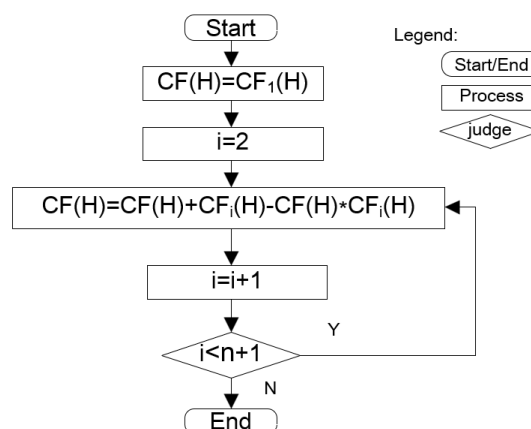


Fig. 2. Calculation process of  $CF(H)$  when multiple rules are matched

To clarify the process, an example is given when a component with unknown type matches all the following three classification and coding rules are matched when reasoning. The value of  $P(H)$  is 0.307, which was calculated by statistical analysis of 8,838 components in the data set.

(1) Rule 1: IF the vertical edge is the longest, and the volume is less than 2 cubic meters, and the number of connecting columns is 0, and the aspect ratio is greater than 0.1 and not less than 0.2, THEN the confidence level that the component is a concrete column. The value of  $P(H|E)$  of Rule 1 is 0.868, which was calculated from the results of rule mining. The value of  $CF(H, E)$  of Rule 1 is 0.868, which was calculated by using Equation (2). The value of  $CF_1(H)$  of Rule 1 is 0.810, which was calculated by using Equation (1).

(2) Rule 2: IF the surface area is less than 10 square meters, the vertical edge is the longest, and the short-to-long ratio is greater than or equal to 0.1 but not less than 0.2, THEN the component is a concrete column. The value of  $CF_2(H)$  of Rule 2 is 0.808 by the same approach.

(3) Rule 3: IF the vertical edge is the longest, and the volume is less than 2 cubic meters, and the short-to-long ratio is greater than or equal to 0.1 and not less than 0.2, THEN the component is considered as a concrete column. The value of  $CF_3(H)$  of Rule 3 is 0.795 by the same approach.

After 2 iterations, the calculation result of the final  $CF(H)$  is 0.993, which was calculated according to Equation (3).

As demonstrated above, the rationality of certainty factor calculation and reasoning can be proven through rigorous mathematical derivation. The result of certainty factor depends on both rules themselves and the matching result. Among which, certainty factor of each rule can be obtained through rule-mining, then, the final certainty factor of one result corresponding to many rules can be calculated by using above equation. The introduction of the credibility reasoning approach ensures the completeness, consistency, and rationality of the reasoning process.

## **7. Case study and validation**

To verify the effectiveness of the proposed method in this study, a customized plugin on the Autodesk Revit was developed. Then, a typical structural BIM model of an engineering project was selected to verify the effectiveness. The architecture of the customized plugin is shown in Fig. 3, which consists of three layers: the data layer, application layer, and interface layer.

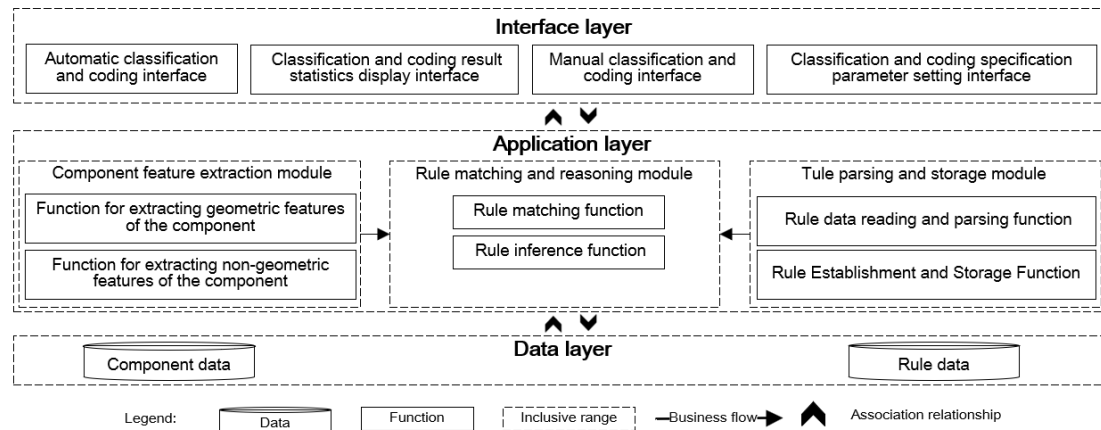


Fig. 3. System architecture for automatic classification and coding of structural components in BIM models

(1) The data layer consists of component data and rule data. The rule data is established and stored in text form, while the component attribute data in the Revit model is obtained through the customization of Autodesk Revit.

(2) The application layer consists of three modules that implement the functionality of BIM model structural component classification and coding, including the component feature extraction module, rule parsing and storage module, and rule matching and reasoning module. The rule parsing and storage module is used to parse the rule data in the data layer, read the rule text file, and convert it into a format that can be understood by the computer. The rule matching and reasoning module processes the parsed rule data and component data, performs reasoning, and inputs the classification and coding results of the components to be classified.

(3) The interface layer provides operational entry points for software users, including the automatic classification and coding interface, the classification and coding result statistics display interface, the manual classification and coding interface, and the classification and coding specification parameter setting interface.

Importing BIM models with incomplete classification and coding attribute into the software allows for automatic identification of unclassified components and addition of coding information. Additionally, the software can export reports on the classification and coding of structural components in BIM models. The user can execute the procedure by following the subsequent steps. First, set classification and coding parameters, including selected standards, the profession of the model to be classified, and the location of the exported report. Second, upon launching the automatic classification and coding program, the system will load the backend rule sets and execute reasoning to add attribute to these components automatically in the current view. Any components that were not able to be automatically classified will be isolated and displayed in the current view, as shown in Fig. 4. Third, the user can view the classification and coding results through the statistics entry. Double-clicking on any component will navigate the system to the location of the component for the user to view automatically. The user can also add attribute manually to the component through the manual classification and coding entry.

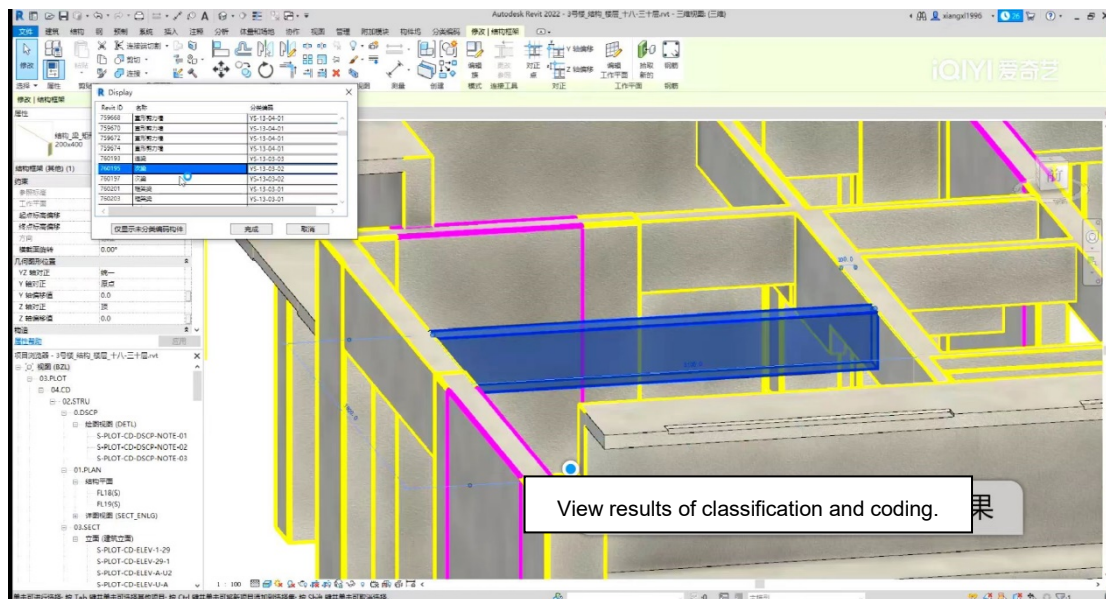


Fig. 4. Example of automatic classification and coding of structural components in BIM models

This study involved importing the structural BIM model of a particular real project into the verification process, which includes 454 components. Among these, 421 components were classified and coded successfully, with 92.7% precision was achieved. These components identified include YS-13-04-01 rectangular shear walls, YS-13-03-03 beams, YS-13-01-04 eaves, and others. The exported component classification and coding report can display the entire list of component codes and other information.

## 8. Discussion and Conclusion

This paper established a method to classify and code structural components of BIM models automatically considering the uncertainty of rules established by using data-driven methods. Furthermore, a customized plugin was developed for validation by using a batch of BIM models from structural design, and 92.7% precision is achieved in the test case. This method helps alleviate the problem of high dependence on domain expert knowledge in rule-based methods, and improves the efficiency and accuracy for replenishing the component attribute. By enabling fast and accurate component classification and coding, the use of BIM technology can be promoted in various stages and tasks such as design, construction, and maintenance. Nevertheless, there is still a risk of overfitting the training model in rule mining, due to the high quality and quantity requirements for the labelling of the structural BIM model training data set. Our future research will further expand on the data mining approach, with focusing on reducing the requirements for labelled training data or exploring the possibility of not labelled training data at all.

## Acknowledgements

This research was sponsored by Country Garden Holdings Company Limited, Guangdong, China.

## References

- [1] David Bryde, Marti Broquetas, Jtirgen Marc Volm, "The project benefits of building information modelling," *International Journal of Project Management*, pp. 971-980, 2013, doi.org/10.1016/j.jiproman.2012.12.001.
- [2] Rebekka Volka, Thu Huong Luu, Johannes Sebastian Mueller-Roemer, Neyir Sevilmis, Frank Schultmann, "Deconstruction project planning of existing buildings based on automated acquisition and reconstruction of building information," *Automation in Construction*, pp. 226-245, 2018, doi:10.1016/j.autcon.2018.03.017.
- [3] GB/T 51269: "Standard for classification and coding of building information model," China Architecture and Architecture Press, 2017.
- [4] Zhai Yiyang, Zhang Chi, Wang Bo, Wang Xue, Liu Kai, Tang Zhongze, "The standardization method and application of the BIM model for interchanges," *Applied Sciences*, pp. 87-89 2022, doi: 10.3390/ap12178787.
- [5] Xuehan Xiong, Antonio Adan, Burcu Akinci, Daniel Huber, "Automatic creation of semantically rich 3D building models from laser scanner data," *Automation in Construction*, pp. 325-337, 2013, doi: 10.1016/j.autcon2012.10.006.
- [6] Pingbo Tang, Burcu Akinci, "Formalization of workflows for extracting bridge surveying goals from laser-scanned data," *Automation in Construction*, pp. 306-319, 2012, doi: 10.1016/i.autcon.2012.09.006.
- [7] Steven L. Brunton, J. Nathan Kutz, "Data-Driven Science and Engineering Machine Learning, Dynamical Systems," and Control, Cambridge University Press, 2022, ISBN: 978-1-009-11563-6.
- [8] Wei Shuangfeng, Liu Minglei, Zhao Jianghong, Huang Shuai, "A survey of methods for detecting indoor navigation elements from point clouds," *Geomatics and Information Science of Wuhan University*, pp. 2003-2011, 2018 doi: 10.13203/j.whugis20180144.
- [9] F. Bosche, C.T. Haas, "Automated retrieval of 3D cad model objects in construction range images," *Automation in Construction*, pp. 499-512, 2008, doi: 10.1016/j.autcon.2007.09.001.
- [10] Tanya Bloch, Rafael Sacks, "Comparing machine learning and rule-based inferencing for semantic enrichment of BIM models," *Automation in Construction*, pp. 256-272, 2018, doi: 10.1016/j.autcon.2018.03.018.
- [11] Koo, B., Shin, B., "Applying novelty detection to identify model element to IFC class misclassifications on architectural and infrastructure building information models," *Journal of Computational Design and Engineering*, pp. 391-400, 2018, doi: 10.1016/j.jcde.2018.03.002.
- [12] Max Ferguson, Seongwoon Jeong, Kincho H. Law, "Worksite object characterization for automatically updating building information models," *Computing in Civil Engineering*, pp. 303-311, 2019, doi: 10.1061/9780784482421.039.
- [13] Tarcisio Mendes de Farias, Ana Roxin, Christophe Nicolle, "A rule-based system for semantical enrichment of building information exchange," *HAL Open Science*, 2014, doi: 10.1016/j.eswa.2015.02.029.
- [14] Michael Belsky, Rafael Sacks, "Semantic enrichment for building information modeling," *Computer-Aided Civil and Infrastructure*, pp. 261-274, 2016, doi: 10.1111/mice.12128.
- [15] Rafael Sacks, Ling Ma, Raz Yosef, Andre Borrmann, Simon Daum, Uri Kattel, "Semantic enrichment for building information modeling: Procedure for compiling inference rules and operators for complex geometry," *Journal of Computing in Civil Engineering*, 2018, doi: 10.1061/(asce)cp.1943-5487.0000705.
- [16] Jin Wu, Jiansong Zhang, "Automated BIM object classification to support BIM interoperability," *Construction Research Congress*, pp. 706-715, 2018, doi: 10.1061/(asce)cp.1943-5487.0000858.
- [17] Liu Ruochen, Mu Caihong, Jiao Licheng, Liu Fang, Cheng Puhua, "Introduction to artificial intelligence," Tsinghua University Press. 2021, ISBN 978-7-302-58468-1.