Proceedings of the 11th International Conference on Applied Informatics
Eger, Hungary, January 29–31, 2020, published at http://ceur-ws.org

Classification Refinement With Category Hierarchy

Judit Tamás, Zsolt Tóth

Eszterházy Károly University
Eger, Hungary
tamas.judit@uni-eszterhazy.hu
toth.zsolt@uni-eszterhazy.hu

Abstract

The concept of classification refinement using hierarchical grouping of categories is presented in this paper. Hierarchical grouping can be determined by heuristic, or existing hierarchical clustering algorithms can be applied to generate tree structures. The concept presented requires the classifier, the grouping of the categories and a threshold value as parameters. The concept is defined to be used for multiple classification tasks. The presented concept can improve the accuracy of classifiers in the case of low confidence level.

Keywords: classification, hierarchical clustering, Miskolc IIS Hybrid Data Set

1. Introduction

These days people depend on technology, our life has become unimaginable without high-tech tools and gadgets. We highly rely on navigation, which give us turn-by-turn directions, traffic congestion information, and alternative routes to a given location. The demand arisen to use navigation in complex buildings like airports, railway stations or hospitals. However, classic Global Positioning Systems do not work in indoor spaces. As a result, Indoor Positioning Systems (IPS) are introduced.

Indoor Positioning Systems can be used to determine the position of people or objects in buildings and closed areas. IPS has been considered as an active research field since the early 1990s, and these systems are detailed in the following surveys [3, 6]. The existing indoor positioning solutions rely on different technologies such

as Infrared [17], ultrasonic [18], magnetic field [9], mobile communication [16], LED [5] or other radio frequency [8, 19, 20] signals.

Indoor positioning is challenging due to the unique properties of the indoor environment. Developers have to make trade-offs between accuracy and cost when they choose a technology. Currently, indoor positioning is vital for smart environments. However, a sufficiently precise, easily accessible, and sustainable industrial standard has not been created yet.

Symbolic positions can be considered as a category, thus the symbolic positioning can be converted into a classification problem. Some well-known classifier accept classes as prediction based on the confidence values. There are some cases when the confidence for each class is relatively small. Hence, the accuracy of these classifiers can vary in a moderate range.

For indoor positioning purposes, a new approach can be introduced. It should increase the accuracy of the classification, and consider the topology of the indoor space.

2. Enhanced classification concept

To boost the performance of these classifiers, a hierarchical grouping of class categories can be introduced. Using hierarchical clustering information of symbolic positions, the accuracy of symbolic indoor positioning algorithms can be improved in case of a low confidence level.

The concept of enhanced classification requires parameters, namely the classifier, the threshold and the dendrogram. The classifier is a method for supervised learning based on the training set and data set, where the target is a discrete attribute. The threshold is a real value between 0 and 1, which determines whether the prediction is accepted or the proposed concept is used. If the confidence value of the predicted class is equal to or higher than the threshold, the classifier method returns with the class. The dendrogram can be predefined by a linkage matrix or it is produced by linkage [1] and distance methods parameters from the topology information.

The tree structure generated by the hierarchical clustering can be seen in Figure 1. The leaf nodes are the rooms, while the root node is the whole described environment.

The tree structure had been modified to include additional information using Python language. The representation of the dendrogram is created with treelib, which enables the traversal in the tree. The identifier of each node is derived from the dendrogram. Each node contains pointers for its parent and its child nodes. The nodes contain a data object, which contains two information. The first information is the universally unique identifier (uuid), which is used for searching purposes. The second is the set of the contained zones, which will be returned as a result by the process.

Based on the improved tree structure, the following process of the enhancement concept is performed.

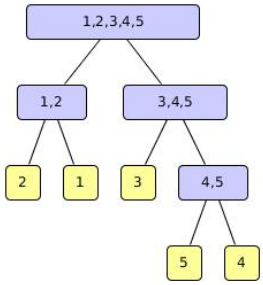


Figure 1: Concept base structure

1. The prediction is performed with the classifier.
2. If the confidence of the predicted class is equal to or higher than the threshold, the process terminates by returning the class as the result.
3. The leaf node in the tree is located using the uuid.
4. Until the confidence of the current node is not reaching the threshold or the root node is reached.
   (a) The parent of this node is selected for examination.
   (b) Its confidence is calculated as the sum of the confidence values of its descendant leaf nodes.
5. The process terminating by returning the contained zones of the lastly examined node.

2.1. Test

In the experiment, the k–NN and the Naive Bayes classifiers are used to the available functionality to return the class probabilities. These classifiers are instance-based classifier, which does not require retraining in case of new instances. The k–NNW denotes the weighted vote version of the k–NN classifier in this paper. The threshold is noted as TH, and $TH \in \{0.6, 0.7, 0.8, 0.9, 1\}$. In this experiment, each linkage method is performed for each classifier and threshold. The linkage methods in the experiment are average, complete, single and weighted. The distance function is selected to be the dissimilarity value of gravitational force-based approach [10, 11, 13]. The gravitational force-based approach is defined in our previous work, it is designed to be used for indoor positioning. The environment is narrowed to rooms on the same level for understandable examination. Different cases can be found in the test, which can present the benefit of the presented concept.

2.1.1. Environment

The Miskolc IIS Hybrid IPS Data set [7, 15] was used to perform the classification. The data set had been recorded in the Miskolc IIS Building of the University of Miskolc using the ILONA System [12, 14, 21]. Each measurement consists of three part, namely the measurement information, the position information and the measurements. The ID and the timestamp of the measurements is stored as the measurement information. Both absolute position with x, y, z coordinates, and symbolic position with uuid and name is saved for each measurement. Sensor information from WiFi, Bluetooth and Magnetometer are included in the measurements. The sensor information is the features for the classification process, while the uuid is the target. These information will be included in the classification process. The topology of the building had been described using IndoorGML [2, 4], which is used to generate the dendrograms. Both the data set and the IndoorGML document uses the same identification for the zones.

To narrow the scope of the experiment, the environment is chosen to be the second floor of the Miskolc IIS Building. Hence the used data set is also narrowed to 431 measurements. From the narrowed data set, the training and the test set are constructed by using stratified sampling with 0.9 and 0.1 ratio. The training and the test sets are fixed during the test. The environment contains 20 zones, and it can be seen in Figure 2. It can represent a general building with narrow corridors, a huge room, which is a lecture hall in this environment, and small office rooms.

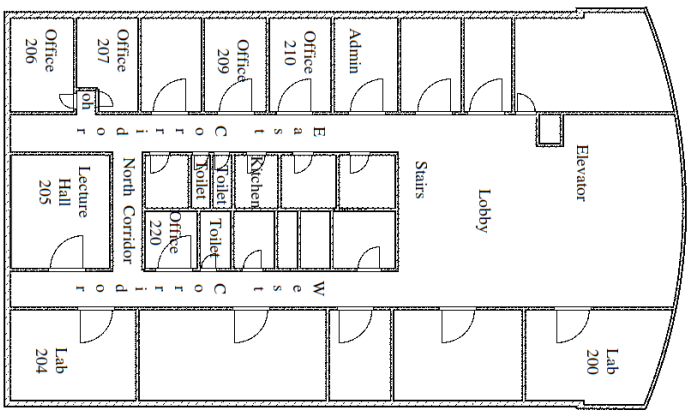


Figure 2: Second floor of the Miskolc IIS Building

However, the Miskolc IIS Hybrid Dataset contains measurements taken in only 5 of these rooms, namely the *East Corridor*, *West Corridor* and *North Corridor*, the *Lobby* and the *Lecture Hall 205*.

2.1.2. Case

To verify the usability of the presented concept, a beneficial case scenario is presented. Although there are cases, where the enhancing concept is not required or applied. For example, 1–NN will always result in 1 confidence during the prediction.

Based on the environment, the weighted linkage method and the gravitational force-based distance, the hierarchical clustering resulted the dendrogram shown in Figure 3.

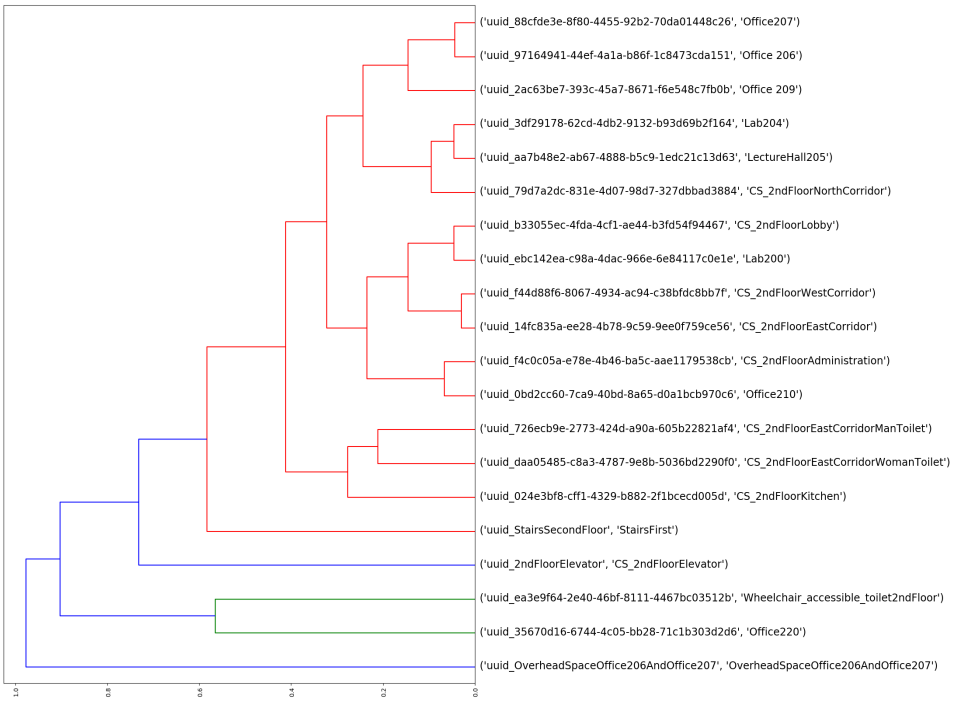


Figure 3: Dendrogram generated by using Gravitational force-based distance and weighted linkage method

The 9-NN classifier was used without a weighted vote to predict the class using the measured values. Based on the dendrogram presented in Figure 3, a tree can be constructed as seen in Figure 4. In this tree, the leaf nodes presented in the dataset have probability values for the given measurement. But only two of these nodes have a non-zero value. The first is marked with 11, and it represents the *East Corridor* room. This room has a 0.328 probability in the classification. The second is the *Lobby* denoted by the number 14 with 0.672 probability. The actual class node is *East Corridor* marked by green background colour on the Figure. A traditional classifier would return with the *Lobby*, because it has the highest probability value.

However, the concept presented in Section 2, instead of returning the predicted

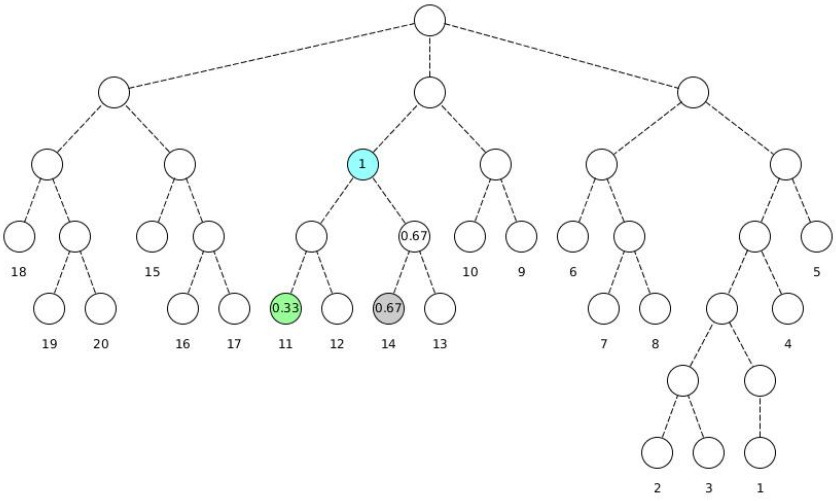


Figure 4: Example case for advantage of enhancement

class, check whether the probability of the predicted class reaches the given threshold. With 0.7 or above threshold, the enhanced classifier locates the predicted class in the tree, and it examines its parent. Hence the parent is not the root node, the process continues. As the predicted node has only one sibling with zero probability, the parent also has the probability value below the threshold. For this reason, the search moves up one level to the parent. The sum of the probabilities of each descendant leaf node is 1, which could pass any threshold. Thus, the last examined node, with the blue background, is the terminating node, which returns the list of its descendant leaf nodes. The result of the classification process consists of only 4 rooms, namely *East Corridor*, *West Corridor*, *Lab200* and *Lobby*. As it can be seen in Figure 4, the actual class if the descendant of the terminating node. Thus the enhanced concept correctly classified the measurement using 4 rooms. However, it could prevent an incorrect classification, which was the goal of the concept.

2.2. Results

The results are stored in a `csv` file for further processing, the schema can be seen in Table 1. Moreover, the file name contains meta-information about the setup, namely the classifier, the linkage method and the threshold.

Correct Classification	Confidence	Set Size	Actual ID	Predicted IDs

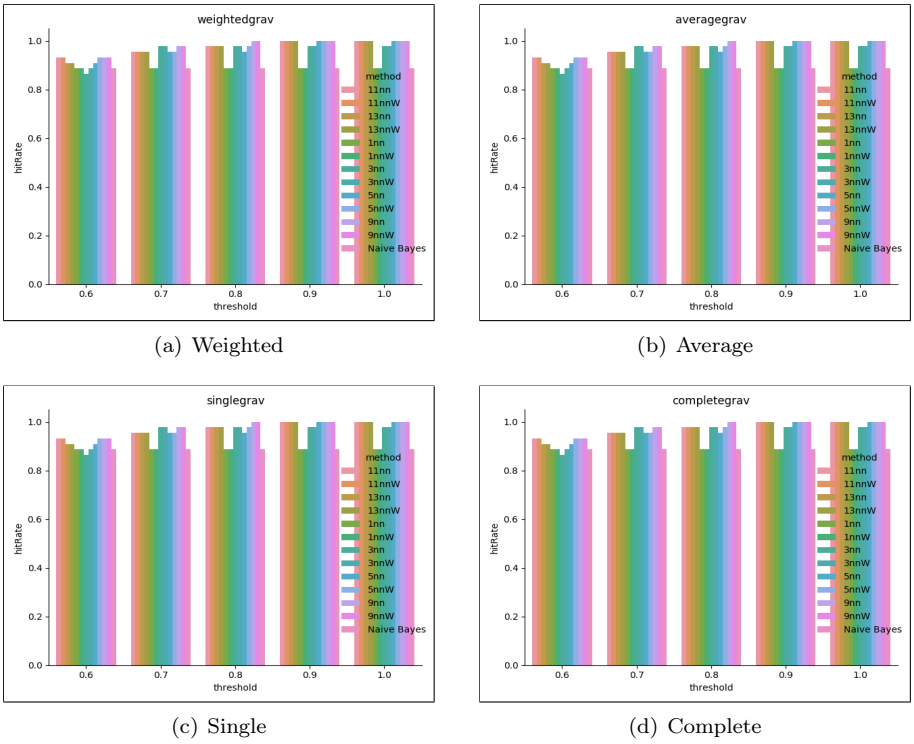
Table 1: Classification results schema

`Correct Classification` can be True or False based on the containment of

the `Actual ID` in the `Predicted IDs` set. `Confidence` is a real value between the threshold and 1, including both value, which represents the accepted confidence of the result. The cardinality of the `Predicted IDs` is stored in the `Set Size` column. The transformation of the selected properties is required for comparison.

2.2.1. Hit

Hit is the associated value for the True or False of `Correct Classification`. Derived from this property of the results, `hitRate` can be calculated for a setup. It is the rate of the correctly classified cases and all the cases to represent the accuracy. Hence, the `hitRate` is a real number in the $[0, 1]$ interval. The goal function is to maximize the `hitRate`.



(a) Weighted

(b) Average

(c) Single

(d) Complete

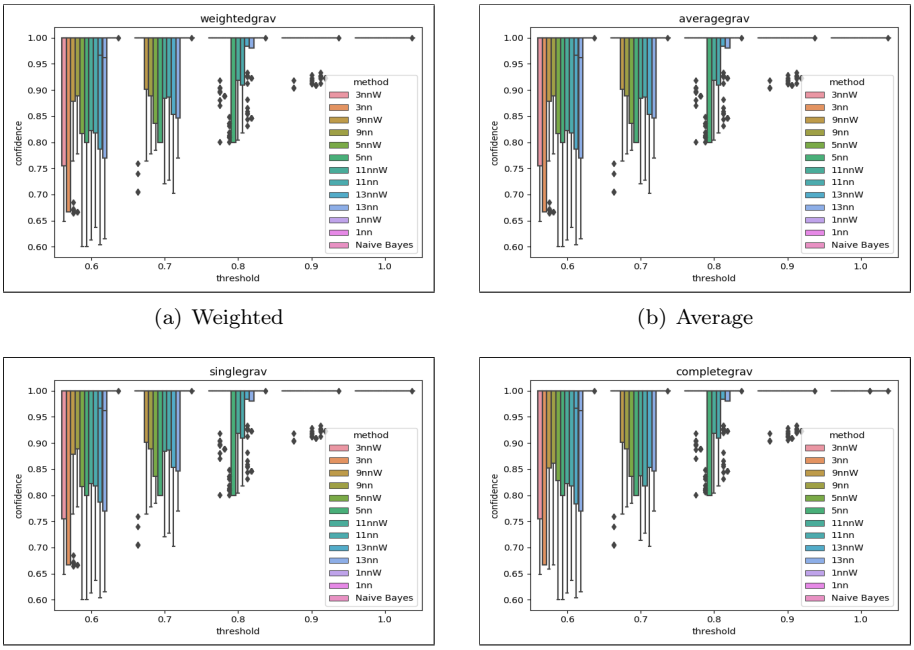
Figure 5: Hit rates of classifiers tested

The `hitRate` values can be seen in Figure 5 for each classifier tested. The values are grouped by both linkage method and threshold. As can be seen, the linkage method does not have a high impact on the `hitRate` in this test. The Figure shows, that 1 `hitRate` was not achieved using a 0.6 or a 0.7 threshold. With 0.8 threshold, the 9–NN and 9–NNW were the few classifiers to achieve 1. Moreover, the set of fully correct classifiers does not differ using 0.9 or 1 as threshold. 1–NN

1–NNW and Naive Bayes classifiers did not use the enhancement in the experiment. Although, 3–NN and 3–NNW were able to increase the `hitRate`, these methods stuck below 1.

2.2.2. Confidence

The confidence property of the results is presented in Figure 6. It is displayed by box plot, grouped by classifier, linkage method and threshold. The goal function is to maximize the confidence values.



(a) Weighted

(b) Average

(c) Single

(d) Complete

Figure 6: Confidences of classifiers tested

As seen in Figure 6, the linkage method has a slight impact on the confidence values. Weighted, average, and single linkage methods resulted in the same statistics of the result set in terms of the confidence property. Compared to the other linkage methods, the complete linkage method has a few hardly noticeable differences. For example, the minimum confidence values using 9–NN and 9–NNW has decreased in case of 0.6 threshold compared to the others. In this setup, the first quartile is also decreased, while there is no outlier detected. However, the 5–NNW and the 13–NNW developed a higher first quartile with the complete linkage method, while outlier is not detected. With 0.7 threshold, 11–NN and 11–NNW achieved a considerably lower first quartile using the complete linkage method, and

the minimum of the 11–NNW slightly decreased. In the rest of the thresholds, the difference lies only in the outlier data.

In terms of the classifiers, it can be said that besides the obvious 1–NN and 1–NNW confidence values, the Naive Bayes resulted also 1 confidence with only one outlier, which is only rounded to 1. The third quartile and the maximum value are 1 regardless of the classifier, the linkage method and the threshold. 9–NN and 9–NNW achieved the notably higher first quartile and minimum using 0.6 threshold. It can be also observed, that the 3–NN and 3–NNW has the first quartile in the 1 value with a 0.7 threshold. However, with 0.8 threshold, 5–NN achieved the equality of minimum and first quartile, while there are no outlier data. Some classifier resulted the first quartile as 1, however, the number of outlier fairly increased. Most classifier has all of their box plot values as 1 using 0.9 threshold, however the number of outlier is still relevant.

2.2.3. Abstraction

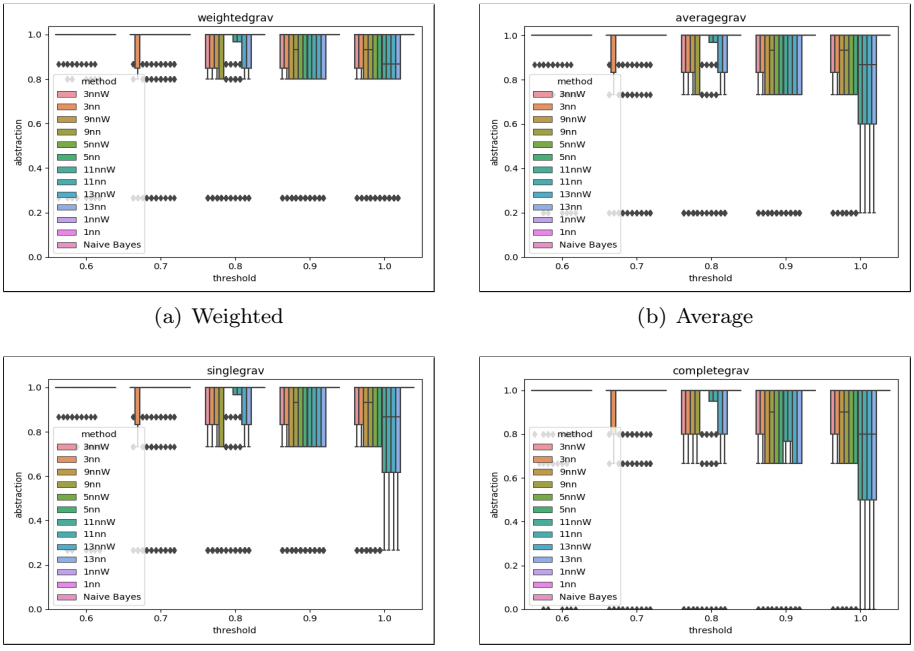
To minimize the size of the resulted list, the abstraction feature is introduced. However, to be consistent with the goal functions of the hitRate and the confidence, the goal for the abstraction should also be maximization. To eliminate the number of rooms from the property, the level of abstraction is designed to be a real number in the $[0, 1]$ range.

$$\hat{a} = 1 - \frac{a - 1}{n - 1} \tag{2.1}$$

Equation 2.1 shows the calculation of abstraction level based on the set size, where a is the set size, n is the number of classes and $\hat{a}$ is the normalized abstraction level. In case the set size is 1, the abstraction level is 1, while the highest possible set size results in 0 as abstraction level.

Figure 7 shows the abstraction levels of classifier, linkage method and threshold setups. As can be seen, linkage method has a high impact on the abstraction feature. From the point of view of minimal abstraction value, the complete linkage method behaves diverse. It shows that some classifiers have cases when the list of all rooms is the prediction results. The weighted linkage method only treats cases as outlier below 0.8 abstraction with every threshold tested. Moreover, compared to the others, the weighted linkage method does not let the minimum abstraction below 0.8, even with a 1 threshold. However, average and single linkage methods mainly differ in the minimal level of abstraction.

In the point of view of the classifiers, the 1–NN, 1–NNW and Naive Bayes have a constant abstraction level with 1. However, using 0.6 as the threshold, other classifiers behave alike, except those have outlier. Only the 3–NN has a minimum lower than 1 in case of 0.7 threshold regardless of the linkage method. With 0.8 threshold, the classifiers that have not lowered their minimum are 5–NN and 5–NNW. Moreover, the amount of change in the case of 11-NN and 11–NNW are also low. The other classifiers took the minimum value to the second row of an outlier in the Figure. Using 0.9 threshold, most of the classifiers took the minimum and



(a) Weighted

(b) Average

(c) Single

(d) Complete

Figure 7: Abstraction of classifiers tested

first quartile values to the second row of an outlier. With the increment of the threshold to 1, 11–NN, 11–NNW, 13–NN and 13–NNW dropped their minimum value last row of outlier, except with weighted linkage.

2.2.4. Discussion

The increment of the threshold does not necessarily improve the classification properties in every case. There is a value, which in case of further increment, does not have any effect, or even reduces the property value. For example, the abstraction is the most reasonable in case of 0.8 or 0.9 as a threshold.

The 3–NNW classifier seems to be the best candidate in the perspective of confidence and abstraction using at least 0.7 as a threshold. Naive Bayes classifier was tested on this environment, however, none of its cases used the concept. Therefore the examination in larger scope is admissible.

The variety of linkage methods does not have an impact on the hit rates, and has a low effect on the confidence property. However, the level of abstraction highly depends on this parameter. For example, the complete linkage method resulted all of the available rooms in some cases, which resulted in a 0 abstraction level. While the weighted, average and single linkage resulted in at least 0.2 abstraction.

In the points of view of the properties, the following can be noticed. When

the accuracy is the main goal, the concept can return all of the rooms as the result, producing a low abstraction level. Moreover, when the level of abstraction is aimed to be as low as possible, the performance of the classification can be poor. For example, Figure 7 shows that the level of abstraction is the best using a 0.6 threshold, the confidence of the classifiers, shown in Figure 6, is weak, and the accuracy is below potential values. Therefore, the threshold and the linkage cannot be based on only one of these features. Hence, the tuning of these properties is required to be examined.

3. Summary

A concept of enhanced classification is presented in this paper. To boost the performance, this concept using hierarchical grouping of class categories. The concept requires the classifier, the threshold and the dendrogram as parameters. The concept is presented with a scenario, which shows its usability. Then the concept is tested in a narrow environment. In the test, the k–NN and Naive Bayes classifiers are selected. The dendrogram is generated by using hierarchical clustering with the dissimilarity value of gravitational force-based approach and weighted, average, single, and complete linkage methods. The results are evaluated using hitRate, confidence, and abstraction properties. However, the properties are conflicting, hence the tuning of these properties is suggested to be further investigated.

Acknowledgements. The first author's research was supported by the grant EFOP-3.6.1-16-2016-00001 ("Complex improvement of research capacities and services at Eszterhazy Karoly University").

References

[1] Blashfield, R. K., Aldenderfer, M. S.: *The literature on cluster analysis*, Multivariate Behavioral Research 13.3 (1978), pp. 271–295.

[2] Ilku, K., Tamas, J.: *IndoorGML Modeling: A Case Study*, in: Carpathian Control Conference (ICCC), 2018 19th International, IEEE, 2018, pp. 633–638.

[3] Koyuncu, H., Yang, S. H.: *A survey of indoor positioning and object locating systems*, IJCSNS International Journal of Computer Science and Network Security 10.5 (2010), pp. 121–128.

[4] Lee, J., Li, K.-J., Zlatanova, S., et al.: *OGC® indoorgml*, Open Geospatial Consortium standard (2014).

[5] Li, L., Hu, P., Peng, C., Shen, G., Zhao, F.: *Epsilon: A Visible Light Based Positioning System.* In: NSDI, 2014, pp. 331–343.

[6] Liu, H., Darabi, H., Banerjee, P., Liu, J.: *Survey of wireless indoor positioning techniques and systems*, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 37.6 (2007), pp. 1067–1080.

[7] *Miskolc IIS Hybrid IPS Data Set*, [Online; Date donated 04-July-2016], URL: http://archive.ics.uci.edu/ml/datasets/Miskolc+IIS+Hybrid+IPS.

[8] Ni, L. M., Liu, Y., Lau, Y. C., Patil, A. P.: *LANDMARC: indoor location sensing using active RFID*, Wireless networks 10.6 (2004), pp. 701–710.

[9] Särkkä, S., Tolvanen, V., Kannala, J., Rahtu, E.: *Adaptive Kalman filtering and smoothing for gravitation tracking in mobile systems* (Oct. 2015), pp. 1–7.

[10] Tamas, J., Toth, Z.: *Topology-Based Evaluation for Symbolic Indoor Positioning Algorithms*, IEEE Transactions on Industry Applications 55.6 (Nov. 2019), pp. 6324–6331, ISSN: 1939-9367, DOI: 10.1109/TIA.2019.2928489.

[11] Tamas, J.: *Hierarchical Clustering based on IndoorGML Document*, in: 2019 IEEE 15th International Scientific Conference on Informatics (INFORMATICS 2019), IEEE, 2019, pp. 411–416.

[12] Tamas, J., Toth, Z.: *Limitation of CRISP accuracy for evaluation of room-level indoor positioning methods*, in: 2018 IEEE International Conference on Future IoT Technologies (Future IoT), Jan. 2018, pp. 1–6, DOI: 10.1109/FIOT.2018.8325585.

[13] Tamas, J., Toth, Z.: *Topology-based Classification Error Calculation for Symbolic Indoor Positioning*, in: Carpathian Control Conference (ICCC), 2018 19th International, IEEE, 2018, pp. 643–648.

[14] Toth, Z.: *ILONA: indoor localization and navigation system*, Journal of Location Based Services 10.4 (2016), pp. 285–302, DOI: 10.1080/17489725.2017.1283453.

[15] Toth, Z., Tamas, J.: *Miskolc IIS hybrid IPS: Dataset for hybrid indoor positioning*, in: 2016 26th International Conference Radioelektronika (RADIOELEKTRONIKA), IEEE, Kosice, Slovakia, Apr. 2016, pp. 408–412.

[16] Wang, S., Green, M., Malkawa, M.: *E-911 location standards and location commercial services*, in: Emerging Technologies Symposium: Broadband, Wireless Internet Access, 2000 IEEE, IEEE, Richardson,TX, USA, Apr. 2000, 5–pp.

[17] Want, R., Hopper, A.: *Active badges and personal interactive computing objects*, Consumer Electronics, IEEE Transactions on 38.1 (1992), pp. 10–20.

[18] Ward, A., Jones, A., Hopper, A.: *A new location technique for the active office*, Personal Communications, IEEE 4.5 (1997), pp. 42–47.

[19] Weissman, Z.: *Indoor location*, White paper, Tadlys Ltd (2004).

[20] Youssef, M., Agrawala, A.: *The Horus WLAN location determination system*, in: Proceedings of the 3rd international conference on Mobile systems, applications, and services, ACM, Seattle, WA, USA, June 2005, pp. 205–218.

[21] Zsolt, T., Péter, M., Richárd, N., Judit, T.: *Data Model for Hybrid Indoor Positioning Systems*, Production Systems and Information Engineering 7 (2015), pp. 67–80.