Conference on
Information Technology and
Data Science

Tuning of Category Hierarchy Enhanced Classification Based Indoor Positioning

Judit Tamás*, Zsolt Tóth

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

Proceedings of the 1st Conference on Information Technology and Data Science
Debrecen, Hungary, November 6–8, 2020
published at http://ceur-ws.org

Abstract

The tuning of classification refinement using hierarchical grouping of categories is presented in this paper. The refinement can improve the accuracy of classifiers in the case of low confidence level and it uses a classifier, a threshold and a dendrogram as parameters. For the examination, the k–NN and the Naive Bayes classifiers are used and the dendrogram will be generated by using linkage method and dissimilarity value of gravitational force-based approach on the topology information. The topology of the environment is described by IndoorGML (Indoor Geographic Markup Language) document. The data set for the classification is part of the Miskolc IIS (Institute of Information Science) Hybrid IPS (Indoor Positioning System) Data set recorded with the ILONA (Indoor Localization and Navigation) System. Three properties are examined of a setup, namely hitRate, confidence and abstraction, however, they are conflicting. A fitness function is introduced using these properties for the purpose of tuning. In this paper, the different weight tuples are examined in the given test environment. The goal of the paper is to examine the weighting possibilities of the hitRate, confidence, and abstraction level features for indoor positioning purposes.

Keywords: Classification, hierarchical clustering

*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").

1. Introduction

These days people dependent on technology, our life has become unimaginable without high-tech tools and gadgets. We highly rely on navigation, which gives 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 [18], ultrasonic [19], magnetic field [9], mobile communication [17], LED [5] or other radio frequency [8, 20, 21] 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 symbolic indoor positioning purposes, a classification refinement using hierarchical grouping of categories had been proposed [12]. Three properties can be established on the proposed method examined, namely hitRate, confidence and abstraction. However, these properties are conflicting, for example, the increment of the hitRate property stimulates the method to return all of the rooms as the result, producing a low abstraction level. Tuning is required to find the balance of these properties to improve the enhancement of the classification based indoor positioning. The goal of the paper is to examine the weighting possibilities of the hitRate, confidence, and abstraction level features for indoor positioning purposes.

2. Enhanced Classification Concept

To boost the performance of the classification, a hierarchical grouping of class categories was introduced [12]. 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 denoted by the uuid, while the root node is the whole described environment.

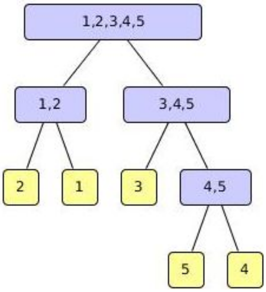

Figure 1. Concept base structure.

The following process of the enhancement concept is performed.

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.

3. Test and Environment

The concept of enhanced classification requires parameters, namely the classifier, the threshold and the dendrogram. 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, well-known and easy to parameterize. 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 the experiment the dendrograms are generated by using linkage methods and dissimilarity value of gravitational force-based approach [10, 11, 14] on the topology information. The linkage methods in the experiment are average, complete, single and weighted, and each linkage method is performed for each classifier and threshold value. The gravitational force-based approach is defined in our previous work, it is designed to be used for indoor positioning.

The Miskolc IIS Hybrid IPS Data Set [7, 16] 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 [13, 15, 22]. Each measurement is composed by three parts, namely the measurement information, the position information and the measured sensor values. The ID and the timestamp of the measurements is stored as the measurement information. Position information part contains both absolute position with x, y, z coordinates, and symbolic position with uuid and name pairs. Sensor information from WiFi, Bluetooth and Magnetometer are included in the measurements. For the classification process, the measured sensor information is the features, while the uuid of the symbolic position is the target.

The topology of the building had been described using IndoorGML [2, 4], which is used to generate the dendrograms. IndoorGML is a standard defined by the Open Geospatial Consortium (OGC) [4], and it represents the indoor spaces as non-overlapping closed objects. The indoor spaces are bounded by physical or fictional boundaries. For each indoor space, the identifier is chosen to be derived from the corresponding space of Miskolc IIS Hybrid Data set.

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. 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.*

Three properties are examined of a setup, namely hitRate, confidence and abstraction. 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`.

Confidence is a real value between the threshold and 1, including both value, which represents the accepted confidence of the result. The goal function is to maximize the confidence values.

To minimize the size of the resulted list, the abstraction feature is introduced.

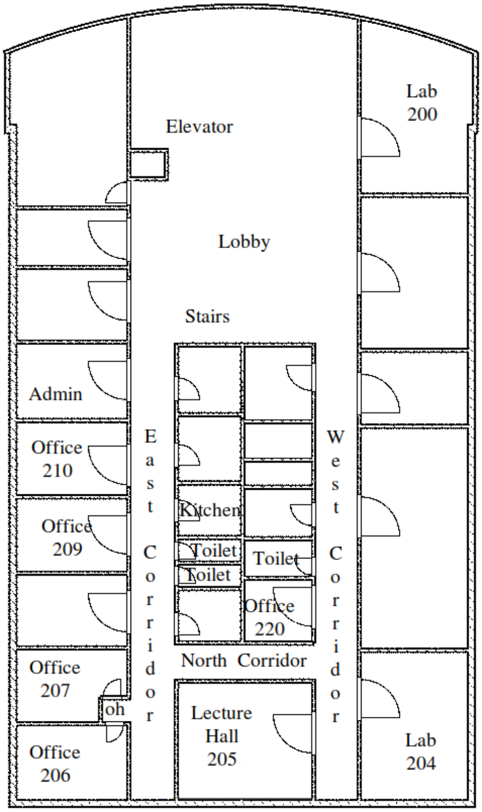

Figure 2. Second floor of the Miskolc IIS Building.

However, to be consistent with the goal functions of the hitRate and the confidence, the goal for the abstraction should also be a 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. Equation (3.1) 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.

$$\hat{a} = 1 - \frac{a-1}{n-1} \tag{3.1}$$

However, when the increment of the hitRate is focused on, the method can return all of the rooms as the result, producing a low abstraction level. In addition, when higher confidence values are aimed at increased threshold, the abstraction level can decrease. Therefore, the goal of the method cannot be based on only one of these properties. Tuning is required to find the balance of these properties to improve the enhancement of the classification based indoor positioning.

A fitness function is introduced using these properties for the purpose of tuning. The introduced fitness function assigns a non-negative weight to each property, where the sum of the weights is 1. The goal of the fitness function is to be maximized.

$$\text{fitness} = w_h \cdot \text{hitRate} + w_c \cdot \overline{\text{confidence}} + w_a \cdot \overline{\text{abstraction}}$$

4. Results

The results are stored in a `csv` file for further processing, the schema can be seen in Table 1. The result contains 1688 rows, where the method, the TH, the linkage method and the weights define a setup.

Table 1. Classification results schema.

Method	TH	Hit	Abstraction	Confidence	w_h	w_a	w_c	Fitness

Among the 1688 setup cases, 756 cases resulted in the highest fitness value in the experiment, and the focus is on these setups. The statistics of the three properties for each classifier can be seen in Table 2.

Table 2. Best Fitness Setups.

method	Average Hit	$\overline{\text{Confidence}}$	$\overline{\text{Threshold}}$	**Count**
1nn	0.89	1	0.8	**180**
1nnW	0.89	1	0.8	**180**
5nn	1	1	0.95	**72**
5nnW	1	1	0.95	**72**
9nn	1	1	0.95	**72**
9nnW	1	1	1	**36**
11nn	1	1	1	**36**
11nnW	1	1	1	**36**
13nn	1	1	1	**36**
13nnW	1	1	1	**36**
Total Result	**0.95**	**1**	**0.89**	**756**

As it can be seen in Table 2, the $1nn$ and the $1nnW$ classifiers are the most frequent, while setups using the Naive Bayes classifier is not presented. The $5nn$, the $5nnW$ and the $9nn$ classifiers are presented mostly after the $1nn$. The average hitRate is 0.95, the average confidence is 1 and the average threshold is 0.89, and these values are not affected by the used linkage method. However, the average abstraction varied with different linkage method, which is shown in Table 3.

Table 3. Average abstraction.

	average	complete	single	weighted	**Total**
1nn	1.00	1.00	1.00	1.00	**1.00**
1nnW	1.00	1.00	1.00	1.00	**1.00**
5nn	0.78	0.74	0.80	0.82	**0.78**
5nnW	0.78	0.74	0.80	0.82	**0.78**
9nn	0.75	0.71	0.77	0.78	**0.75**
9nnW	0.75	0.71	0.77	0.78	**0.75**
11nn	0.73	0.68	0.75	0.77	**0.73**
11nnW	0.73	0.68	0.75	0.77	**0.73**
13nn	0.73	0.67	0.75	0.77	**0.73**
13nnW	0.73	0.67	0.75	0.77	**0.73**
Total	**0.87**	**0.85**	**0.88**	**0.89**	**0.87**

As Table 3 shows, the average abstraction of 1*nn* and 1*nnW* classifiers are obviously 1, while the second best value is in the case of 5*nn* and 5*nnW* with 0.78. In the point of view of the linkage method, the weighted linkage method resulted in 0.89 average abstraction, while the last in the order is complete linkage with 0.85. The overall average abstraction of the highest fitness valued cases is 0.87.

The distribution of the average hit among the best cases according to the threshold values can be seen in Table 4.

Table 4. Best cases- Average Hit Based on Threshold.

	0,6	0,7	0,8	0,9	1,0
1nn	0,9	0,9	0,9	0,9	0,9
1nnW	0,9	0,9	0,9	0,9	0,9
5nn				1,0	1,0
5nnW				1,0	1,0
9nn				1,0	1,0
9nnW					1,0
11nn					1,0
11nnW					1,0
13nn					1,0
13nnW					1,0
Total Result	0,9	0,9	0,9	1,0	1,0

It can be seen from the data in Table 4 that until 0.9 threshold, classifiers could not reach the highest presented fitness value with the exception of 1*nn* and

$1nnW$. With 0.9 threshold, the $5nn$, the $5nnW$ and the $9nn$ classifiers could reach 1 average hit value. With the 1 threshold, only the $1nn$ and the $1nnW$ classifier could not reach 1 average hit value.

In the point of view of the weights, the statistic made of the cases with the best presented fitness value can be illustrated in Table 5.

Table 5. Statistic of weights.

	Hit weight			Confidence weight			Abstraction weight		
	AVG	Min	Max	AVG	Min	Max	AVG	Min	Max
1nn	0	0	0	0.5	0.1	0.9	0.5	0.1	0.9
1nnW	0	0	0	0.5	0.1	0.9	0.5	0.1	0.9
5nn	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
5nnW	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
9nn	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
9nnW	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
11nn	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
11nnW	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
13nn	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
13nnW	0.5	0.1	0.9	0.5	0.1	0.9	0	0	0
Total	**0.26**	**0**	**0.9**	**0.5**	**0.1**	**0.9**	**0.24**	**0**	**0.9**

From Table 5 we can see that the average weight of the hit and the abstraction are similar with 0.26 and 0.24 value, while the average weight of confidence is the double with 0.5. While both hit and abstraction could be eliminated from the fitness value calculation in some cases, the weight of the confidence is at least 0.1. In the case of $1nn$ and $1nnW$, the hit is completely eliminated, while in the other classifiers resulted in the best fitness value presented eliminated the abstraction property.

The fitness value of the most frequent weights is presented according to the threshold and the classifier using single linkage can be seen in Figure 3. The weight for the hit and the confidence is 0.5, while the abstraction is eliminated.

As shown in Figure 3, using 0.6 threshold, the Naive Bayes, the $1nn$ and the $1nnW$ classifiers have the highest fitness value. When the threshold is increased by 0.1, the $3nn$, the $3nnW$ take the lead, and $9nn$ and $9nnW$ classifiers are also surpass the previous fitness value. However, by further increasing the threshold, the $1nn$, the $1nnW$, the $3nn$, the $3nnW$ and the Naive Bayes classifier could not reach the highest fitness value presented. The other classifiers have increment in their fitness value while the threshold is increased. With 0.9 threshold, the $5nn$, the $5nnW$, and $9nn$ classifiers could reach the highest fitness value, while the other classifiers could reach this fitness value using 1 as threshold.

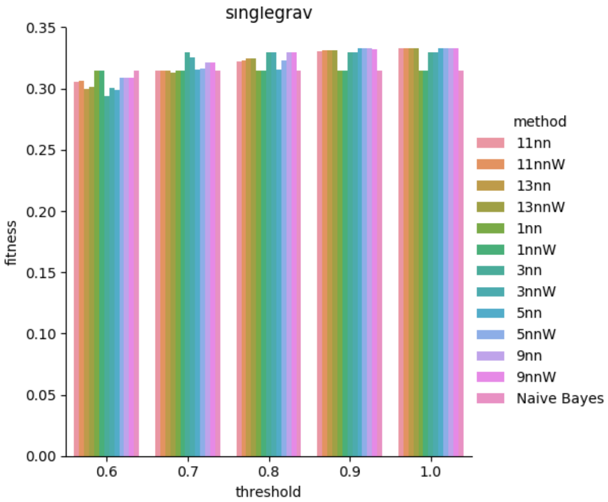

Figure 3. Single linkage with most frequent weights.

4.1. Discussion

We can make observation based on three point-of-view, namely the classifiers, the threshold and the weights.

The $1nn$ and the $1nnW$ classifiers occurred most frequently among the cases with the best fitness value presented. Contrary to the Naive Bayes classifier, which could not reach this value with any of its setups. However, the $5nn$, the $5nnW$ and the $9nn$ classifiers were the second most frequent in the narrowed result set.

With given weights, the $5nn$, the $5nnW$ and the $9nn$ could reach the highest fitness value with 0.9 threshold. With lower threshold, there were cases when the $1nn$ and the $1nnW$ classifiers could reach the highest fitness value, however the average hit rate in this cases is 0.9. Using 1 as threshold, every other classifier presented in the best fitness valued setups reached 1 as average hit.

In most of the cases, only two of the three property is considered when calculating the fitness value. The $1nn$ and the $1nnW$ classifiers neglect the hit property, while the other classifiers neglect the abstraction property. However, the weights of other two properties are equals, thus they are equally important.

5. Conclusion

The tuning of classification refinement using hierarchical grouping of categories is presented in this paper. For the examination, the k–NN and the Naive Bayes classifiers were used and the dendrogram was generated by using linkage method and dissimilarity value of gravitational force-based approach on the topology information. A linear fitness function was introduced using these properties for the purpose of tuning.

The investigation of the fitness function shows that instead of three properties, the setups with the highest fitness value neglect one of the properties. The other two properties were proved equally important in the cases. However, a tested classifier could not reach the highest fitness value with any of its setups. This research has thrown up many questions in need of further investigation.

In the future, the category hierarchy enhanced classification based indoor positioning concept is planned to be examined in two ways. First is the expansion of the test environment in three dimension, which helps to test the concept in multi-floored environment. The second way is the modification of the fitness function by using non-linear elements.

References

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

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

[3] H. Koyuncu, S. H. Yang: *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] J. Lee, K.-J. Li, S. Zlatanova, T. Kolbe, C. Nagel, T. Becker: *OGC® indoorgml*, Open Geospatial Consortium standard (2014).

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

[6] H. Liu, H. Darabi, P. Banerjee, J. Liu: *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*, http://archive.ics.uci.edu/ml/datasets/Miskolc+IIS+Hybrid+IPS, [Online; Date donated 04-July-2016].

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

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

[10] J. Tamas, Z. Toth: *Topology-based Evaluation for Symbolic Indoor Positioning Algorithms*, IEEE Transactions on Industry Applications (2019), pp. 1–1, issn: 0093-9994,
doi: 10.1109/TIA.2019.2928489.

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

[12] J. Tamas, Z. Toth: *Classification Refinement with Category Hierarchy*, in: The 11th International Conference on Applied Informatics (ICAI 2020), published at http://ceur-ws.org, 2020, pp. 358–369.

[13] J. Tamas, Z. Toth: *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.

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

[15] Z. Toth: *ILONA: indoor localization and navigation system*, Journal of Location Based Services 10.4 (2016), pp. 285–302,
doi: 10.1080/17489725.2017.1283453, eprint: http://dx.doi.org/10.1080/17489725.2017.1283453,
url: http://dx.doi.org/10.1080/17489725.2017.1283453.

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

[17] S. Wang, M. Green, M. Malkawa: *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.

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

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

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

[21] M. Youssef, A. Agrawala: *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.

[22] T. Zsolt, M. Peter, N. Richard, T. Judit: *Data Model for Hybrid Indoor Positioning Systems*, PRODUCTION SYSTEMS AND INFORMATION ENGINEERING 7 (2015), pp. 67–80.