

Short thesis for the degree of
doctor of philosophy (PhD)

Classification based
Symbolic Indoor Positioning

by Judit Kunné Tamás
Supervisor: Zsolt Tóth, PhD



UNIVERSITY OF DEBRECEN
Doctoral School of Informatics
Debrecen, 2021

Contents

1	Introduction	3
1.1	Research Goals	5
2	Results	6
2.1	Symbolic Indoor Positioning	6
2.1.1	Miskolc IIS Hybrid Data Set	7
2.1.2	Benchmarking	9
2.2	Topology-based Evaluation	12
2.2.1	Gravitational-force based approach	13
2.2.2	Capacity and Distance function	14
2.2.3	Experiments	15
2.2.4	Comparison	18
2.3	Enhanced Classification	21
2.3.1	Concept	24
2.3.2	Proof of concept	26

2.3.3	Results	28
2.3.4	Real-life scenario	30
Bibliography		31
3.	Bevezetés	37
3.1.	Kutatási célok	39
4.	Eredmények	41
4.1.	Szimbolikus Beltéri Helyzetmeghatározás . .	41
4.1.1.	Miskolc IIS Hibrid Adathalmaz . . .	42
4.1.2.	Teljesítményértékelés	44
4.2.	Topológia Alapú Kiértékelés	47
4.2.1.	Tömegvonzás alapú megközelítés . .	49
4.2.2.	Kapacitás és Távolság függvény . . .	50
4.2.3.	Kísérletek	51
4.2.4.	Összehasonlítás	54
4.3.	Javított Osztályozás	57
4.3.1.	Elgondolás	61
4.3.2.	Koncepció igazolása	63
4.3.3.	Eredmények	65
4.3.4.	Valós életbeli kísérlet	67
Irodalomjegyzék		69

1. Introduction

Indoor Positioning Systems (IPS) have been considered as an active research field since the early 1990s, however, the topic gained popularity in the 2010's with the widespread of smartphones and it is still a hot topic these days. Indoor positioning systems are used to determine the position of people or objects in buildings and closed areas. The smart environment can be enhanced with indoor positioning and navigation services. With Industry 4.0 [5], the phenomenon of Smart Factories [3, 8, 4] has emerged, where the location awareness is significant. Indoor Positioning Systems are usually classified by the used technologies and the type of location.

Proximity based, Absolute and Symbolic positions [2] can be determined by Indoor Positioning Systems. The selection of the position type of indoor positioning system highly depends on the application area. Proximity-based

positioning can be suitable for applications where not the exact location, but closeness to a well-known place or object, like Beacon, is required. Absolute position is given by coordinates and it is used by highly accurate and robust systems. Symbolic positioning determine the location of an object as a well-defined part of a building.

Symbolic positions can be considered as a category label, thus the symbolic positioning can be converted into a classification problem. Classification is a well-studied part of data mining, therefore numerous well-known classification methods could be applied to the indoor positioning. The evaluation of the symbolic indoor positioning methods is based on the CRISP approach. The CRISP approach does not differentiate between the wrongly predicted classes, hence every misclassification is equally wrong. Besides CRISP logic, application-specific approaches are usually used for the evaluation of classifiers.

Application-specific methods use some domain knowledge. For symbolic indoor positioning, a part of the domain knowledge can be incorporated into the description of the environment. To the best of our knowledge, there is no application-specific classification evaluation method for symbolic indoor positioning.

1.1 Research Goals

In my research, the following three goals were set. The first goal was to create a data set, which can be used to compare solutions for symbolic indoor positioning purposes. It should contain measurements from multiple sensors. Based on the created data set, various classifiers planned to be tested. To the best of our knowledge, there was no other such data set available. The second goal is to create an application-specific approach for classification error calculation for indoor positioning purposes. It should incorporate the domain knowledge, the topology, to the calculation. The usability of the proposed method needs to be examined, and compared to the traditional, CRISP approach. Application-specific evaluation of classification error is usually used in other fields, but as far as we know, this kind of method was not available for symbolic indoor positioning. The third goal is to develop a novel method to improve the classification accuracy for symbolic indoor positioning purposes using environment topology as domain knowledge. Most of the classification-based solutions do not use domain knowledge. Applying this knowledge, we

can provide a more error-tolerant method.

2. Results

2.1 Symbolic Indoor Positioning

Thesis 1. *Room-level indoor positioning can be considered as a classification problem. I created a data set, which allows benchmarking of classification-based symbolic indoor positioning methods.*

Data sets are widely used for evaluation and comparison of various machine learning algorithms. The UCI Machine Learning Repository [7] contains more than 300 data sets for various tasks such as classification, clustering and regression. This repository contains two data sets [10, 11] related to indoor positioning and the measurements were performed in a multi-building and multi-floor environment. However these data sets allow the comparison of

various indoor positioning methods, but they are limited to one technology.

2.1.1 Miskolc IIS Hybrid Data Set

The Miskolc IIS (Institute of Information Science) Hybrid IPS Dataset was constructed to provide a static context for the evaluation of the classification-based symbolic indoor positioning methods. This dataset is available in the UCI Machine Learning Repository. It contains measurements of multiple built-in sensors of mobile phones, such as Bluetooth, WiFi and Magnetometer. The construction of the data set was made at the end of February, 2016. The measurements were recorded in a three-story building of the University of Miskolc using the ILONA System [12].

The measurements were performed in the Institute of Information Science at the University of Miskolc. The data set cover about 50% of the building because access to the offices and storage rooms is restricted. The measurements were taken in 21 symbolic positions in the building, called Zones, hence 21 classes are distinguished. The data set contains 1539 measurements, 32 WiFi Access Points, and 20 Bluetooth devices. The measured reference points can be seen in Figure 2.1, where each colour represents a zone.

In order to facilitate the data mining, evaluation, and processing tasks, the database of measurements was ex-

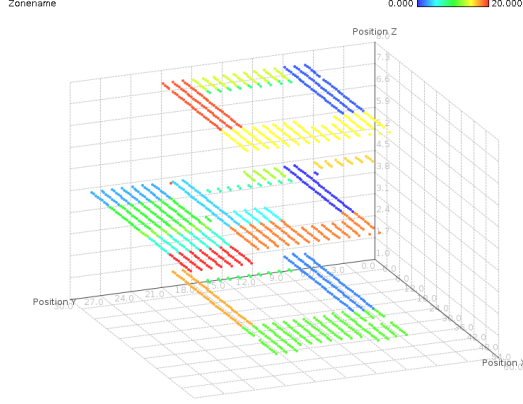


Figure 2.1: The Reference Points of the Data Set

ported into CSV (Comma-Separated Values) format, as seen in Figure 4.2. The header contains the following fields: the *ID* of the measurement, *timestamp*, the absolute coordinates called *x*, *y*, *z*, the *name* of the symbolic position, the *Zone ID* of the symbolic position and the magnetic field coordinates denoted with *magnetometerX*, *magnetometerY*, *magnetometerZ*.

The list of the WiFi Access Points is fixed for each measurement, and the Access Points are presented by their SSID. Fields indexed between 11–42 represent the RSSI value of the corresponding Access Point. In the data set,

null denotes the missing values in the WiFi RSSI fields.

Each position between 43 and 65 has a corresponding Bluetooth device that is identified by a string, which contains its name and MAC address. The record contains 1 if the device was within range and 0 otherwise.

Measurement Information		Position Information		Measurements		
ID	Timestamp	Absolute	Symbolic	Magnetometer	WiFi APs	Bluetooth
1	2	3-5	6-7	8-10	11-42	43-65

Figure 2.2: CSV Structure of the Data Set

2.1.2 Benchmarking

The Miskolc IIS Hybrid Indoor Positioning System Data set was used as the data set during the evaluation process. The following classification methods were tested: k -NN, Naive Bayes, Decision Tree, ANN, and Rule induction. The selection criterion of the tested classifiers was the representation of both model-based and instance-based classifiers, including the most popular methods. The evaluation process was performed with RapidMiner with the exception of the Multilayer Perceptron (MLP) because its topology was optimized with a genetic algorithm of Weka framework. The Artificial Neural Network is denoted as ANN, the k -NNW represents the weighted vote variant of

the k -NN. Decision tree with C4.5 algorithm is denoted as Decision Tree, while with the ID3 algorithm is denoted by ID3 according to the RapidMiner component name.

The main factor during the evaluation of each model was the classification accuracy of the CRISP approach based on the validation samples. The accuracy is the proportion of correctly predicted cases. It provides us with more certainty that an unknown measurement will be classified correctly. The CRISP approach does not differ among the predicted misclassified symbolic positions, hence it considered each other symbolic position equally wrong.

The data set is partitioned into the training and the validation subsets. The records of each set contain attributes related to the sensors and the id of the symbolic position as category label. Due to the small data set, the training and validation sets were built by the stratified sampling of the data set with a 0.9 and 0.1 ratio.

Figure 2.3 shows the accuracies achieved during the experiments with the selected classification methods. Simulations show that Rule Induction and Decision Tree classifiers are not suggested due to their low accuracy. The Naive Bayes achieved an acceptable accuracy, while the k -NN and ANN managed to earn the best results. Among the methods tested in Rapid Miner, the k -NN classifier had the best performance when the k parameter was 3 and uses weighted major vote, as is shown in Figure 2.3. The

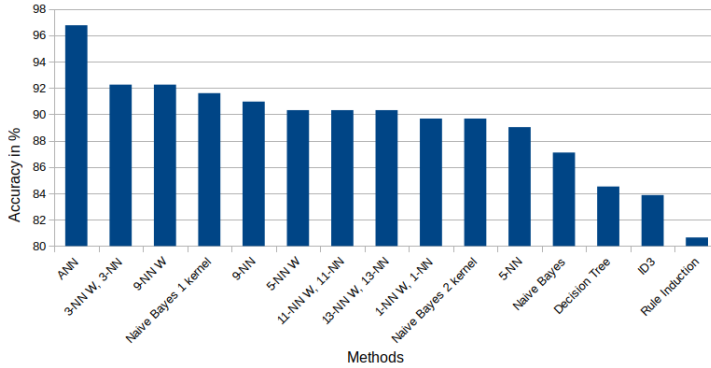


Figure 2.3: Performance of the k -NN, Rule Induction, Decision Tree, Artificial Neural Network and Naive Bayes

Artificial Neural Network could perform 96.77% accuracy with 0.9 learning rate, 0.5 momentum, 380 training time and 1 hidden layer with 30 neurons. From the instance-based classifiers, the k -NN method seems to be the best classifier. In the case of model-based classifiers, the ANN achieved outstanding results among them. Hence, both an instance-based and a model-based classifier could be applied efficiently to indoor positioning purposes.

2.2 Topology-based Evaluation

Thesis 2. *I proposed a topology-based evaluation method for classification-based indoor positioning algorithms which allows a more detailed evaluation.*

In a further examination of experimental results produced by symbolic indoor positioning methods, a remarkable behaviour can be detected. A more accurate classification method predicts further the symbolic positions from the original location when it is misclassified, than a less accurate classifier. Furthermore, less accurate classifiers often predict the neighboring, and nearby symbolic position. In conclusion, a different approach is recommended to be considered as an alternative of CRISP for symbolic indoor positioning purposes.

Topology-based evaluation of symbolic indoor positioning methods requires two things: a classification error calculation method and a formal description of the indoor environment as domain knowledge to quantify the classification error. The topology-based approach should measure the similarity of the Zones based on their distance and size. We can establish requirements, which seems to be essential for topology-based classification error calculation for symbolic indoor positioning purposes. Firstly, the error values

should be proportional to the sizes of the Zones. Secondly, the layout of the Zones should have a high impact on the error values. Lastly, the classification error should not be symmetric due to the size differences of the Zones.

2.2.1 Gravitational-force based approach

The gravitational force-based approach was designed to consider the topology in the classification error calculation. Let Z be the finite set of the disjunct rooms. The approach requires the determination of capacity and distance functions. The capacity ($V : Z \rightarrow \mathbb{R}^+$) function maps each room to a positive real value. The distance ($d : Z^2 \rightarrow \mathbb{R}$) function determines how far a room is from another or it measures the dissimilarity of the rooms.

The gravitational force ($F_g : Z^2 \rightarrow \mathbb{R}$) [9] measures the similarity between two rooms. The gravitational force is calculated as seen in Equation 2.1, and it is non-negative and symmetric derived from the symmetry of distance function.

$$F_g(Z_i, Z_j) = \frac{V(Z_i)V(Z_j)}{d(Z_i, Z_j)^2} \quad (2.1)$$

While the gravitational force represents the similarity between rooms, their difference is required for error calculation. The δ ($\delta : Z^2 \rightarrow \mathbb{R}^+$) function is introduced to

represent the difference of two rooms. The δ function can be calculated as seen in (2.2) and it ranges in $]0, 1[$.

$$\delta(Z_i, Z_j) = \frac{1}{1 + F_g(Z_i, Z_j)} \quad (2.2)$$

The classification error should also be proportional to the sizes of the rooms and asymmetric due to the size differences. The ϵ ($\epsilon : Z^2 \rightarrow \mathbb{R}$) function is introduced to measure the classification error while fulfils these requirements. The ϵ is non-symmetric in the $[0, 1]$ range, as can be seen in (2.3).

$$\epsilon(Z_i, Z_j) = \frac{V(Z_i) * \delta(Z_i, Z_j)}{V(Z_i) + V(Z_j)} \quad (2.3)$$

The gravitational force-based approach fulfils the requirement to consider the topology in the classification error calculation.

2.2.2 Capacity and Distance function

The gravitational force-based approach requires the determination of capacity (V) function and distance (d) function. The capacity of a room can be calculated in two or three dimensions. The capacity of a room in two-dimensional space is called the area. In three-dimensional space, the capacity of a room is the volume.

The distance of two rooms can be calculated in a coordinate system or a graph model. Two types of reference points can be distinguished to a room, namely centroid and nearest points. The centroid reference point is the geometric centre of a room, which is a global feature. The nearest points are local reference points, hence the nearest point depends on the placement of the two rooms. It can be either the nearest corner points from a room to another or any points of two parallel line segments. The indoor environment could be mapped to a graph where the nodes are the rooms, and the edges are their connections. In the graph, the distance of two nodes can be measured with their shortest path. The shortest path between two nodes can be found by different algorithms in different graphs.

2.2.3 Experiments

Applicability of the proposed error calculation method is demonstrated with two experiments. Therefore, we set up two experiment cases with different reference points, and with the same distance function and capacity function. With the distance and the area values, the gravitational force, then the δ values can be calculated. Based on the δ values and the areas of the Zones, the error matrix can be constructed.

Simulation

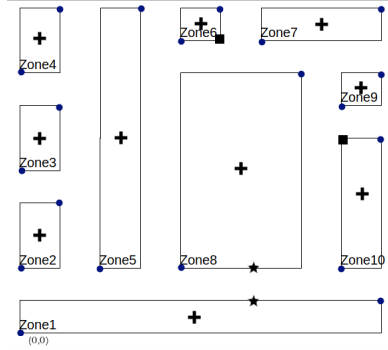


Figure 2.4: The Layout of the Test Environment

The test environment was given in a two-dimensional space. The environment consists of 10 Zones, and we assume that they are rectangular without overlapping, as it can be seen in Figure 2.4.

The Zones are defined by their two diagonally opposite points and marked with dots in Figure 2.4. The coordinates of the Zones are measured in units. In both experiments, the capacity function is selected to be the area of the Zones. The distance of the two Zones is calculated based on the Euclidean distance function of their reference points. The selection of the reference points distinguishes

the two experiments, that are the centroid and the boundary distance cases. The centroid of each Zone are marked with a cross in Figure 2.4 and the nearest boundary points of Z_6 and Z_{10} pair are marked with squares, and for the Z_8, Z_1 Zone pair, they are denoted with stars.

Based on our experimental results, the presented method results proportional error values to the Zone sizes, takes the layout of the Zones into the count and the error values of two Zones are asymmetric. Therefore the presented classification error calculation method considers the topology.

Experiment

To further test the applicability of the proposed gravitational force-based method, the experiment in a real-life environment is performed. IndoorGML (Indoor Geographic Markup Language) [6, 1] is a standard defined by the Open Geospatial Consortium (OGC), which is an open format to describe the topology. This document describes the rooms by two forms in three-dimensional space, one is with vertices and the other is with a bounding box for speed-up purposes. Hence, the IndoorGML provides the possibility of the usage of coordinate-based distance calculation. In addition, the document contains transitions between rooms, which can be mapped to a graph.

The Institute of Information Science (IIS) Building was

modeled with IndoorGML standard. The coordinates are determined based on the available grid, thus the data set and the IndoorGML use the same coordinate system. The three-dimensional lower- and upper corner coordinates are used in distance and capacity calculation.

For distance calculation, the Euclidean distance had been chosen using centroid reference point, and the capacity is the volume. The calculated error values have a very low average, and the standard deviation should be higher, and the highest error value is lower than the half of the possible range. Therefore, the gravitational force-based approach can be an alternative to the CRISP approach in the evaluation of symbolic indoor positioning methods.

2.2.4 Comparison

Experiments were performed in order to compare the CRISP and the proposed topology-based classification evaluation method. Comparison was performed over the Miskolc IIS Hybrid data set and the topology of the building described with IndoorGML. The same classifiers were ranked based on CRISP and three variants of the proposed topology-based method as presented in the first thesis.

Three variant of the proposed topology-based classification evaluation is tested using different distance function. The first variant calculates the distance between the cen-

troids with the Euclidean distance function. The second variant calculates the distance between the nearest points with the Euclidean distance function. The third variant uses Dijkstra’s shortest path algorithm on the generated graph from the IndoorGML.

Results

The tested classifiers were ranked based on the calculated classification error. In the case of exact classification error values, the classifiers are grouped. The largest group in the CRISP evaluation consists of 5 classifiers, with 4 classifiers in the case of graph distance, while both centroid distance and nearest point distance produced the largest group with only 2 classifiers. Rank-based ordering of the classifiers was also analysed. Two kinds of classifiers can be distinguished. The classifiers with fixed positions form the first type, the second group consists of classifiers with diverse positions. Among the classifiers in the highly diverse group, there are ones with significant diversity. In the comparison, classifiers with slightly diverse positions are the most frequent.

The relative performance of the classifiers can be considered as the sensitivity of the evaluation method. For each evaluation method, the classification error values are divided by the best value, which is the lowest error value

of the given evaluation. This relative performance of the classifiers can be seen in Figure 2.5, where the x -axis shows the multiplier, the number of times the classifier has more error, than the best classifier in the evaluation method.

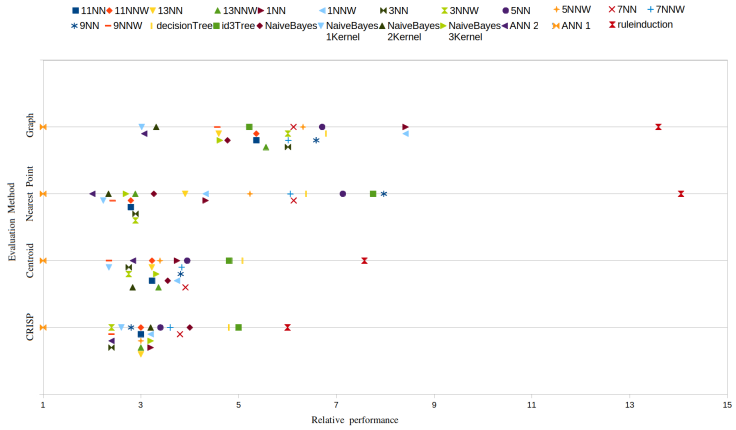


Figure 2.5: Relative Performance of Tested Classifiers in Evaluation Cases

The nearest point distance and the graph distance methods are more alike, while the CRISP and the centroid distance methods also show more similarity. Nearest point distance function method developed the largest range between the best and the worst method. As a general result,

even the poorly performing variant of gravitational force-based approach was able to slightly outperform the results of the CRISP approach.

2.3 Enhanced Classification

Thesis 3. *I designed a classification-based symbolic indoor positioning method enhanced by hierarchical clustering, which considers the topology of the building based on the confidence of the classification.*

Well-known classifiers accept classes as a prediction based on the confidence values. There are some cases when the confidence for each class is relatively small. To boost the performance of these classifiers for symbolic indoor positioning purposes, a hierarchical grouping of class categories can be introduced. Hierarchical clustering algorithms can be applied using topological description to generate the tree.

The room representation is performed using IndoorGML standard. The IndoorGML document is transformed to be used for clustering purposes. The lower and upper corners of the bounding box and all the vertices of the rooms are preserved. Their transitions are not convenient to create

a hierarchy among rooms. After the elimination of unnecessary features, the capacity is introduced as new feature. Furthermore, the identifier and the name are merged for indexing purposes.

The grouping is performed in the agglomerative way using two distance functions, namely the Euclidean and the gravitational-force based approach.. The δ of the gravitational-force based approach can be considered as a virtual distance among the rooms. Euclidean distance is used as a distance metric and volume is used as capacity function for the gravitational force-based approach. The average, centroid, complete, median, single, ward and weighted linkage methods had been tested. Centroid, median and ward linkage methods can not be applied with gravitational force-based distance or any distance function other than the Euclidean distance.

The created cluster models are compared based on the generated dendrograms. Tanglegram is used to compare tree diagrams, which measures the quality of the two dendrogram alignment as entanglement. The entanglement value is the base of the evaluation.

The trees generated by the same distance function but different linkage method are compared to each other, and then optimized to minimize the entanglement value. These entanglement values are not necessary symmetric due to the implemented brute force approach. Experimental re-

sults show, that most of the linkage functions achieved similar average entanglement values. However, the single linkage method shows a significant difference in both distance functions.

Euclidean and Gravitational force-based distance using weighted linkage
with Optimized Entanglement: 9.3803

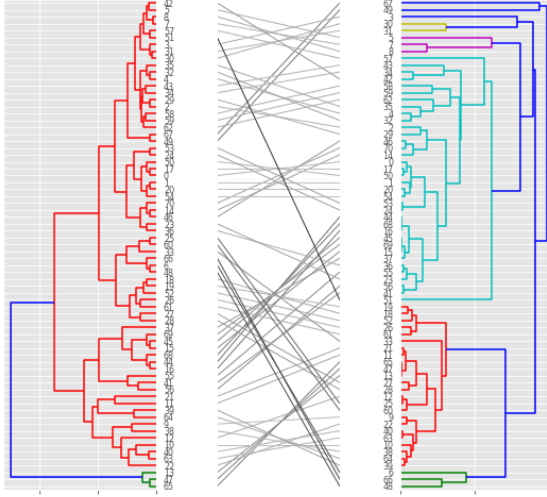


Figure 2.6: Euclidean and Gravitational Force-based Distance using Weighted Linkage Method with 10000 Iterations

In addition, the two distance function can be exam-

ined by using the same linkage function. The weighted linkage had been selected for the comparison due to the compatibility and lower entanglement values. Figure 2.6 shows that the two dendrograms are highly diverse. Euclidean distance can result in similar hierarchical clustering using most of the linkage methods. However, these linkage methods seem understandable to the given purpose, the gravitational force-based distance could reflect the topology, the arrangement of the rooms by the distances in the dendrogram.

2.3.1 Concept

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 a classifier, a threshold and a dendrogram. The threshold is a real value between 0 and 1, which determines whether the prediction is accepted or the proposed concept is used.

The tree structure generated by the hierarchical clustering can be seen in Figure 2.7. 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. Each node contains pointers for its parent and its child nodes. The nodes contain data object,

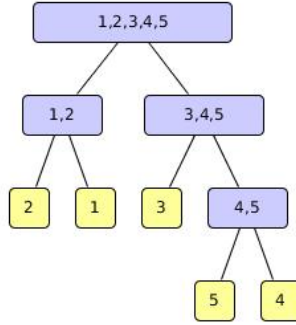


Figure 2.7: Concept Base Structure

which contains three information. It contains the uuid for searching purposes, the set of the contained zones, and the size of this set.

Based on the improved tree 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 by 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.

2.3.2 Proof of concept

The proposed concept is examined in an experiment. 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 selected, because they are easy to parametrize. The threshold is noted as TH , and $TH \in \{0.6, 0.7, 0.8, 0.9, 1\}$. In this experiment, each linkage method tested is performed for each classifier and threshold. The distance function is selected to be the dissimilarity value of the gravitational force-based approach. The narrow the scope of the experiment for understandable examination, the environment is chosen to be the second floor of Miskolc IIS Building defined in IndoorGML. However, the Miskolc IIS Hybrid Dataset contains measurements taken in only 5 of these rooms.

Different cases can be found in the test, which can present the benefit of the presented concept. The 9-NN classifier was used without weighted vote to predict the class using the measured values. Based on the built dendrogram using weighted linkage and gravitational force-based distance, a graph can be constructed as seen in Figure 2.8. Only two of these nodes have a non-zero probability value, marked with 11 and 14. The actual class node is marked by green background colour on the graph. A traditional classifier would return the node with grey background, because it has the highest probability value.

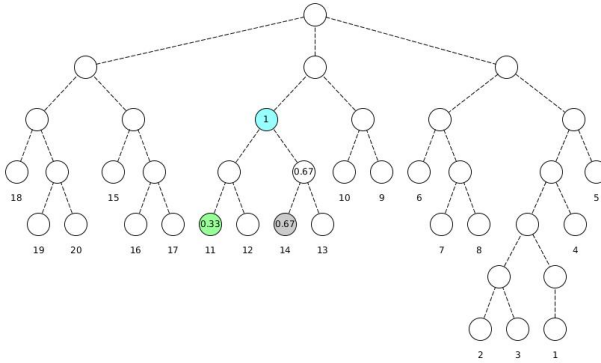


Figure 2.8: Example Case for Advantage of Enhancement

However, the concept instead of returning the predicted

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 graph, and it examines its parent. 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 searching process 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 actual class is the descendant of the terminating node. Thus, the enhanced concept correctly classified the measurement using only 4 rooms.

2.3.3 Results

Three properties are examined of each setup to examine the classification, namely **hitRate**, confidence and abstraction. **hitRate** 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 values, which represents the accepted confidence of the result. The goal function is to maximize the confidence values. To mini-

mize 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.

When the accuracy is the main goal, the concept can return all 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. Therefore, the threshold and the linkage can not be based on only one of these features. To help find the balance, a fitness function is introduced using these features. The introduced fitness function can be seen in Equation 2.4, where w denotes the weight of the given property. The aim is to maximize the fitness value.

$$fitness = w_h \cdot hitRate + w_c \cdot \overline{confidence} + w_a \cdot \overline{abstraction} \quad (2.4)$$

The different linkage methods resulted in noticeable differences in the fitness value. Based on the results, the weighted and the single linkage methods seem to be advisable. Naive Bayes classifier was tested on this environment, however, none of its cases used the concept. Therefore the examination in larger scope is admissible.

2.3.4 Real-life scenario

A real-life scenario is performed to investigate the possibilities of the proposed concept. It is similar to the previous experiment, but the distance function and the environment is different. The distance function is selected to be the commonly-used Euclidean distance, and the dissimilarity value of gravitational force-based approach. The results are also examined in the view-of-point of hitRate, confidence and abstraction level.

We can determine combinations, which perform the best according to multiple criteria. For example, cases with the maximal possible value in the experiment as Max Set Size is eliminated. Another example is, the elimination of cases where the average set size is not in $(1, 10]$ range. Then the data set is examined separately based on the distance function. Based on the results, the concept could increase the accuracy of well-known classifiers, while the number of predicted classes is reasonably modest.

Bibliography

- [1] Ogc indoorgml-with corrigendum.
<http://docs.opengeospatial.org/is/14-005r4/14-005r4.html>. [Online; accessed 13-October-2017].
- [2] Christian Becker and Frank Dürr. On location models for ubiquitous computing. *Personal and Ubiquitous Computing*, 9(1):20–31, 2005.
- [3] Ulises Carrasco, Pedro Daniel Urbina Coronado, Mahmoud Parto, and Thomas Kurfess. Indoor location service in support of a smart manufacturing facility. *Computers in Industry*, 103:132 – 140, 2018.
- [4] Jim Davis, Thomas Edgar, James Porter, John Bernaden, and Michael Sarli. Smart manufacturing, manufacturing intelligence and demand-dynamic per-

- formance. *Computers & Chemical Engineering*, 47:145 – 156, 2012. FOCAPO 2012.
- [5] Heiner Lasi, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann. Industry 4.0. *Business & information systems engineering*, 6(4):239–242, 2014.
- [6] J Lee, Ki-J Li, S Zlatanova, TH Kolbe, C Nagel, and T Becker. Ogc® indoorgml. *Open Geospatial Consortium standard*, 2014.
- [7] M. Lichman. UCI machine learning repository, 2013.
- [8] Shaoping Lu, Chen Xu, Ray Y. Zhong, and Lihui Wang. A rfid-enabled positioning system in automated guided vehicle for smart factories. *Journal of Manufacturing Systems*, 44:179 – 190, 2017.
- [9] Isaac Newton. *The Principia: mathematical principles of natural philosophy*. Univ of California Press, 1999.
- [10] Joaquin Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Joan P Avariento, Tomás J Arnau, Mauri Benedito-Bordonau, and Joaquin Huerta. Uji-indoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localiza-

- tion problems. In *Proceedings of the fifth conference on indoor positioning and indoor navigation*, 2014.
- [11] Joaquin Torres-Sospedra, David Rambla, Raul Montoliu, Oscar Belmonte, and Joaquin Huerta. Ujiindoorloc-mag: A new database for magnetic field-based localization problems. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, pages 1–10. IEEE, 2015.
- [12] Zsolt Toth. Ilona: indoor localization and navigation system. *Journal of Location Based Services*, 10(4):285–302, 2016.



Registry number: DEENK/330/2020.PL
Subject: PhD Publication List

Candidate: Judit Kunné Tamás
Doctoral School: Doctoral School of Informatics
MTMT ID: 10060570

List of publications related to the dissertation

Foreign language scientific articles in Hungarian journals (1)

1. Tóth, Z., Magnucz, P., Németh, R., **Tamás, J.**: Data Model for Hybrid Indoor Positioning Systems.
Prod. Syst. Inf. Eng. 7, 67-80, 2015. ISSN: 1785-1270.

Foreign language scientific articles in international journals (2)

2. **Tamás, J.**, Tóth, Z.: Topology-Based Evaluation for Symbolic Indoor Positioning Algorithms.
IEEE Trans. Ind. Appl. 55 (6), 6324-6331, 2019. ISSN: 0093-9994.
DOI: <http://dx.doi.org/10.1109/TIA.2019.2928489>
IF: 3.488
3. **Tamás, J.**, Tóth, Z.: Classification-based symbolic indoor positioning over the Miskolc IIS Data-set.
J. Locat. Based Serv. 12 (1), 2-18, 2018. ISSN: 1748-9725.
DOI: <http://dx.doi.org/10.1080/17489725.2018.1455992>

Hungarian conference proceedings (1)

4. **Tamás, J.**, Tóth, Z.: Osztályozáson alapuló pozicionálási módszerek vizsgálata a Miskolci Informatikai Épület Hibrid Adathalmaza alapján.
In: Műszaki tudomány az Észak-Kelet Magyarországi régióban 2016 : konferencia előadásai.
Szerk.: Bodzás Sándor, Debreceni Akadémiai Bizottság Műszaki Szakbizottsága, Debrecen, 664-668, 2016. ISBN: 9789637064333

Foreign language conference proceedings (7)

5. **Tamás, J.**, Tóth, Z.: Classification Refinement With Category Hierarchy.
In: Proceedings of the 11th International Conference on Applied Informatics (ICAI 2020). Ed.: by Gergely Kovásznai, István Faze kas, Tibor Tómacs, CEUR, Eger, 358-369, 2020, (CEUR Workshop Proceedings, ISSN 1613-0073 ; 2650.)
6. **Tamás, J.**: Hierarchical Clustering based on IndoorGML Document.
In: 2019 IEEE 15th International Scientific Conference on Informatics : Informatics 2019 Proceedings. Eds.: William Steingartner, Stefan Korecko, Anikó Szakál, IEEE, Piscataway, 411-416, 2019. ISBN: 9781728131788





7. Ilku, K., **Tamás, J.**: IndoorGML modeling: a case study.
In: Proceedings of the 2018 19th International Carpathian Control Conference (ICCC). Eds.:
Dániel Drótos, József Vásárhelyi, László Czap, Ivo Petrás, IEEE, Danvers, 633-638, 2018.
ISBN: 9781538647622
8. **Tamás, J.**, Tóth, Z.: Limitation of CRISP accuracy for evaluation of room-level indoor positioning
methods.
In: Proceedings of 2018 IEEE International Conference on Future IoT Technologies (Future
IoT), IEEE, Washington, 1-6, 2018. ISBN: 9781538612088
9. Ilku, K., **Tamás, J.**: Topology-based Classification Error Calculation based on IndoorGML
Document.
In: The 11th Conference of PhD Students in Computer Science : volume of short papers :
CS2, University of Szeged, Szeged, 101-105, 2018.
10. **Tamás, J.**, Tóth, Z.: Topology-based classification error calculation for symbolic indoor
positioning.
In: Proceedings of the 2018 19th International Carpathian Control Conference (ICCC). Eds.:
Dániel Drótos, József Vásárhelyi, László Czap, Ivo Petrás, IEEE, Danvers, 643-648, 2018.
ISBN: 9781538647622
11. Tóth, Z., **Tamás, J.**: Miskolc IIS hybrid IPS: dataset for hybrid indoor positioning.
In: Proceedings of the 26th International Conference Radioelektronika 2016, IEEE,
Piscataway, 408-412, 2016. ISBN: 9781509016747

Total IF of journals (all publications): 3,488

Total IF of journals (publications related to the dissertation): 3,488

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on
the basis of the Journal Citation Report (Impact Factor) database.

02 November, 2020



Doktori (PhD) értekezés tézisei

Osztályozás alapú
szimbolikus beltéri
helyzetmeghatározás

Kunné Tamás Judit
Témavezető: Dr. Tóth Zsolt



DEBRECENI EGYETEM
Informatikai Tudományok Doktori Iskola
Debrecen, 2021

3. fejezet

Bevezetés

A beltéri helymeghatározó rendszereket az 1990-es évek eleje óta aktív kutatási területnek tekintik, azonban a téma a 2010-es években vált népszerűvé az okostelefonok elterjedésével, és napjainkban is aktuális kutatási téma. A beltéri helymeghatározó rendszereket emberek vagy tárgyak helyzetének meghatározására használják épületekben és zárt területeken. Beltéri helymeghatározó és navigációs szolgáltatásokkal javítható az okos otthon koncepció. Az Ipar 4.0 [5] alkalmazásával megjelent az intelligens gyárak [3, 8, 4] jelensége, ahol fontos szempont a helyismeret. A beltéri helymeghatározó rendszereket általában a használt technológiák és a hely típusa szerint csoportosítják.

A beltéri helymeghatározó rendszerek közelségi, abszolút és szimbolikus pozíciókat [2] határozhatnak meg. A pozíciótípus kiválasztása nagymértékben függ a beltéri helyzetmeghatározás alkalmazási területétől. A közelségen alapuló helyzetmeghatározás alkalmas lehet olyan felhasználásra, ahol nem a pontos hely, hanem egy közismert hely vagy tárgy, például a jeladó közelsége szükséges. Az abszolút helyzetet koordináták adják meg, és nagy pontosságú és robusztus rendszerek használják. A szimbolikus pozíció az épület egy jól körülhatárolható része.

A szimbolikus pozíciókat kategóriacímeként tekinthetjük, így a szimbolikus pozicionálás osztályozási problémának tekinthető. Az osztályozás az adatbányászat egy jól tanulmányozott része, ezért számos jól ismert osztályozási módszert lehetne alkalmazni a beltéri helymeghatározásra. A szimbolikus beltéri pozicionálási módszerek értékelése a CRISP megközelítésen alapszik. A CRISP megközelítés nem tesz különbséget a tévesen előre jelzett osztályok között, ezért minden téves osztályozást egyformán tévesnek tekint. A CRISP logika mellett általában az alkalmazás-specifikus megközelítéseket alkalmazzák az osztályozók kiértékeléséhez.

Az alkalmazás-specifikus módszerek bizonyos tartományismeretet használnak. A szimbolikus beltéri pozicionáláshoz a tartományi ismeretek egy része beépíthető a környezet leírásába. Legjobb tudomásunk szerint a szimbolikus

beltéri pozicionáláshoz nincs alkalmazás-specifikus osztályozási kiértékelési módszer.

3.1. Kutatási célok

Kutatásom során a következő három célt tűztem ki. Az első cél egy olyan adathalmaz létrehozása volt, amely felhasználható szimbolikus beltéri helymeghatározási célú megoldások összehasonlítására. Az adathalmaznak több szenzor mérését kell tartalmaznia. A létrehozott adathalmaz alapján különféle osztályozókat tervezek tesztelni. Legjobb tudomásunk szerint más ilyen adatkészlet nem állt rendelkezésre. A második cél egy alkalmazás-specifikus megközelítés létrehozása a beltéri helymeghatározási célú osztályozási hibák kiszámításához. Be kell építenie a tartományi ismereteket, a topológiát a számításba. Meg kell vizsgálni a javasolt módszer használhatóságát, és össze kell hasonlítani a hagyományos, CRISP megközelítéssel. Az osztályozási hibák alkalmazás-specifikus értékelését általában más területeken alkalmazzák, de tudomásunk szerint ez a fajta módszer nem volt elérhető szimbolikus beltéri pozicionáláshoz. A harmadik cél egy új módszer kifejlesztése az osztályozási pontosság javítására szimbolikus beltéri helymeghatározási célokra, a környezet topológiájának felhasználásával. Az osztályozáson alapuló megoldások többsége

nem használ tartományi ismereteket. Ezen ismeretek alkalmazásával hibátűrőbb módszert nyújthatunk.

4. fejezet

Eredmények

4.1. Szimbolikus Beltéri Helyzetmeghatározás

1. Tézis <i>A szobaszintű beltéri helyzetmeghatározás osztályozási problémának tekinthető. Készítettem egy adathalmazt, amely lehetővé teszi az osztályozás alapú szobaszintű beltéri helyzetmeghatározási módszerek teljesítményértékelését.</i>

Az adathalmazokat széles körben használják a különféle gépi tanulási algoritmusok értékelésére és összehasonlítására. Az UCI Machine Learning Repository [7] több, mint

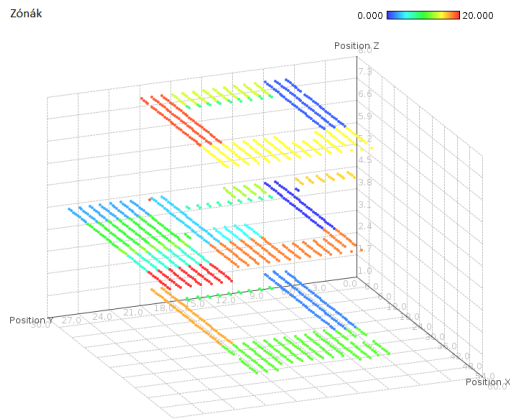
300 adathalmazt tartalmaz különböző feladatokhoz, például osztályozáshoz, klaszterezéshez és regresszióhoz. Ez az adattár két adathalmazt tartalmaz [10, 11], amelyek a beltéri helymeghatározáshoz kapcsolódnak, és a méréseket több épületben és több emeletből álló környezetben hajtották végre. Ezek az adathalmazok lehetővé teszik a különböző beltéri pozicionálási módszerek összehasonlítását, de csak egy technológiára korlátozódnak.

4.1.1. Miskolc IIS Hibrid Adathalmaz

A Miskolc IIS hibrid IPS adathalmaz úgy állítottuk össze, hogy statikus kontextust biztosítson az osztályozáson alapuló szimbolikus beltéri helymeghatározási módszerek kiértékeléséhez. Ez az adathalmaz elérhető az UCI Machine Learning Repository-ban. Tartalmazza a mobiltelefonok több beépített szenzorának - például a Bluetooth, a Wi-Fi és a Magnetométer - méréseit. Az adathalmaz 2016. február végén került rögzítésre. A méréseket a Miskolci Egyetem háromemeletes épületében rögzítettük az ILONA Rendszer [12] segítségével.

A méréseket a Miskolci Egyetem Informatikai Intézetében végeztük. Az adathalmaz az épület körülbelül 50% -át lefedi, mivel az irodákhoz és a raktárakhoz való hozzáférés korlátozott. A méréseket az épület 21 szimbolikus pozíciójában, az úgynevezett zónákban hajtották végre, ezért

4.1. SZIMBOLIKUS BELTÉRI HELYZETMEGHATÁROZÁS⁴³



4.1. ábra. Az Adathalmaz Referencia Pontjai

21 osztályt különböztethetünk meg. Az adathalmaz 1539 mérést, 32 WiFi hozzáférési pontot és 20 Bluetooth eszközt tartalmaz. A mért referenciapontok a 4.1 ábrán láthatók, ahol minden szín egy zónát képvisel.

Az adatbányászat, a kiértékelés és a feldolgozási feladatok megkönnyítése érdekében a mérések adatbázisát CSV (Comma-Separated Values) formátumban exportáltam, amelynek struktúrája a 2.2 ábrán látható. A fejléc a következő mezőket tartalmazza: a mérés azonosítója: *ID*, *időbélyeg*, a *x*, *y*, *z* nevű abszolút koordinátákat, a szimboli-

Mérés Információ		Pozíció Információ		Mért Értékek		
ID	Időbélyeg	Abszolút	Szimbolikus	Magnetométer	WiFi Apk	Bluetooth Eszközök
1	2	3-5	6-7	8-10	11-42	43-65

4.2. ábra. Az Adathalmaz CSV Struktúrája

kus pozíció *neve*, a szimbolikus pozíció *Zóna azonosítója* és a mágneses mező koordinátái *magnetometerX*, *magnetometerY*, *magnetometerZ* néven.

A WiFi hozzáférési pontok listája minden méréshez rögzített, az hozzáférési pontokat pedig az SSID azonosítójuk jelöli. A 11-42 között indexelt mezők a megfelelő hozzáférési pont RSSI-értékét képviselik. Az adathalmazban a *null* jelöli a hiányzó értékeket a WiFi RSSI mezőkben.

Minden pozíció 43 és 65 között egy megfelelő Bluetooth eszköz szerepel, amelyet egy karakterlánc azonosít, amely tartalmazza a nevét és a MAC címét. A rekord 1 értéket tartalmaz, ha az eszköz hatótávolságon belül van, minden más esetben 0 értéket.

4.1.2. Teljesítményértékelés

Az értékelési folyamat során a Miskolc IIS hibrid IPS adathalmaz került felhasználásra. A következő osztályozási módszereket teszteltük: k -NN, Naive Bayes, Döntési fa, ANN és Szabály indukció. A tesztelt osztályozók kiválasztási kri-

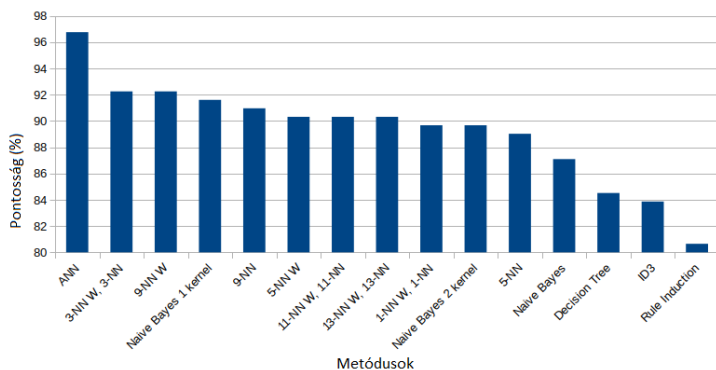
4.1. SZIMBOLIKUS BELTÉRI HELYZETMEGHATÁROZÁS45

tériuma a modellalapú és a példányalapú osztályozók reprezentálása volt, beleértve a legnépszerűbb módszereket is. Az értékelési folyamat a RapidMiner alkalmazásával történt, kivéve a Többrétegű Perceptront, mivel topológiáját a Weka keretrendszer genetikai algoritmusával optimalizáltam. A mesterséges neurális hálózatot ANN-nak jelöljük, a k -NNW a k -NN súlyozott szavazati változatát jelenti. A C4.5 algoritmussal rendelkező döntési fát Decision Tree, míg az ID3 algoritmust az ID3 jelöli a RapidMiner komponens elnevezésének megfelelően.

Az egyes modellek értékelése során a fő tényező a CRISP módszer validálási mintákon alapuló osztályozási pontossága volt. A pontosság a helyesen megjósolt esetek aránya. Ez nagyobb biztonságot nyújt számunkra, hogy egy ismeretlen mérés helyesen lesz osztályozva. A CRISP megközelítés nem tesz különbséget a tévesen osztályozott szimbolikus pozíciók között, ezért egyformán tévesnek tartott az eredetitől eltérő összes többi szimbolikus pozíciót.

Az adathalmaz fel lett osztva tanító és validációs adathalmazokra. Minden halmaz rekordjai tartalmazzák az érzékelőkhöz kapcsolódó attribútumokat és a szimbolikus pozíció azonosítóját címkeként. A kis adathalmaz miatt a tanító és validációs halmazok rétegzett mintavételével segítségével készültek, 0,9 és 0,1 aránnyal.

A 4.3 ábra a kiválasztott osztályozási módszerek kísérlet során elért pontosságait mutatja. A szimulációk azt



4.3. ábra. A k -NN, Szabály Indukció, Döntési Fák, Mesterséges Neurális Háló és Naive Bayes Osztályozók Elért Teljesítményei

mutatják, hogy a szabály indukció és a döntési fa osztályozók alacsony pontosságuk miatt nincsenek javasolva. A Naiv Bayes elfogadható pontosságot ért el, míg a k -NN és ANN osztályozóknak sikerült a legjobb eredményt elérni. A Rapid Minerben tesztelt módszerek közül a k -NN osztályozó akkor érte el a legjobb teljesítményt, amikor a k paraméter 3 volt, és súlyozott szavazatot használt, ami a 4.3 ábrán látható. Az ANN 96,77 pontosságot tud elérni 0,9 tanulási sebességgel, 0,5 momentummal, 380 képzési idővel és 1 rejtett rétegben elhelyezkedő 30 neuronokkal. A példányalapú osztályozók közül a k -NN módszer tűnik a legjobb osztályozónak szobaszintű beltéri helyzetmeghatározás céljára. A modell alapú osztályozók között az ANN kiemelkedő eredményeket ért el. Ennélfogva mind a példányalapú, mind a modellalapú osztályozó hatékonyan alkalmazható beltéri helymeghatározási célokra.

4.2. Topológia Alapú Kiértékelés

2. Tézis *Definiáltam egy topológia-alapú értékelési módszert az osztályozáson alapuló beltéri helyzetmeghatározási algoritmusokhoz, amely részletesebb értékelést tesz lehetővé.*

A szimbolikus beltéri pozicionálási módszerekkel előál-

lított kísérleti eredmények további vizsgálata során figyelemre méltó viselkedés mutatható ki. A pontosabb osztályozási módszer az eredeti pozíciótól távolabbi pozíciót jósol, ha rosszul osztályozzák, mint egy kevésbé pontos osztályozó. Továbbá a kevésbé pontos osztályozók gyakran szomszédos és közeli pozíciót ad eredményül. Összegzőként elmondható, hogy a szimbolikus beltéri helymeghatározás céljából más megközelítést javasolunk a CRISP alternatívájának.

A szimbolikus beltéri pozicionálási módszerek topológiai alapú értékeléséhez két dologra van szükség: egy osztályozási hiba számítási módszerre és a beltéri környezet formális leírására, mint tartományismeretre az osztályozási hiba számszerűsítéséhez. A topológia alapú megközelítésnek a zónák hasonlóságát távolságuk és méretük alapján kell mérnie. Megállapíthatunk követelményeket, amelyek elengedhetetlennek látszanak a topológia alapú osztályozási hiba kiszámításához szimbolikus beltéri helymeghatározási feladatoknál. Először is, a hibaértékeknek arányosnak kell lenniük a zónák méretével. Másodszor, a zónák elrendezésének nagy hatással kell lennie a hibaértékekre. Végül az osztályozási hiba nem lehet szimmetrikus a zónák méretbeli különbségei miatt.

4.2.1. Tömegvonzás alapú megközelítés

A tömegvonzás alapú megközelítés úgy lett tervezve, hogy figyelembe vegye a topológiát az osztályozási hiba kiszámításakor. Legyen Z diszjunkt zónák véges halmaza. A megközelítés megköveteli a kapacitás és a távolság függvényeinek meghatározását. A $(V : Z \rightarrow \mathbb{R}^+)$ kapacitás függvény minden zónához egy pozitív valós értéket rendel. A $(d : Z^2 \rightarrow \mathbb{R})$ távolság függvény határozza meg, hogy egy szoba milyen távolságra van a másiktól, vagy mennyire különböznek.

A tömegvonzás $(F_g : Z^2 \rightarrow \mathbb{R})$ [9] a két szoba közötti hasonlóságot méri.

A tömegvonzás kiszámítása a 4.1 egyenletben látható, nem negatív és szimmetrikus a távolságfüggvény szimmetriájából adódóan.

$$F_g(Z_i, Z_j) = \frac{V(Z_i)V(Z_j)}{d(Z_i, Z_j)^2} \quad (4.1)$$

Míg a tömegvonzás a hasonlóságot képviseli, a hiba kiszámításához az eltérésekre van szükség. Erre került bevezetésre a δ ($\delta : Z^2 \rightarrow \mathbb{R}^+$) függvény. A δ függvény kiszámítása a (4.2) egyenletben látható és a $]0, 1[$ tartományban van.

$$\delta(Z_i, Z_j) = \frac{1}{1 + F_g(Z_i, Z_j)} \quad (4.2)$$

Az osztályozási hibának arányosnak kell lennie a szobák méretével és aszimmetrikusnak kell lennie a méretbeli különbségek miatt. Az ϵ ($\epsilon : Z^2 \rightarrow \mathbb{R}$) függvény az osztályozási hiba mérésére szolgál, miközben megfelel ezeknek a követelményeknek. A ϵ nem szimmetrikus, a $[0, 1]$ tartományban van, amint ez a 4.3 egyenletben is látható.

$$\epsilon(Z_i, Z_j) = \frac{V(Z_i) * \delta(Z_i, Z_j)}{V(Z_i) + V(Z_j)} \quad (4.3)$$

A tömegvonzás alapú megközelítés teljesíti azt a követelményt, hogy az osztályozási hiba kiszámításakor vegye figyelembe a topológiát.

4.2.2. Kapacitás és Távolság függvény

A tömegvonzás alapú megközelítés megköveteli a kapacitás (V) és a távolság (d) függvény meghatározását. A helyiség kapacitása két vagy három dimenzióban számolható. A kétdimenziós térben lévő helyiség kapacitását területnek nevezzük, míg háromdimenziós térben a helyiség kapacitása a térfogat.

Két szoba távolsága koordináta-rendszerben vagy gráf modellen számolható. Minden szobához kétféle referenciapont különböztethető meg, nevezetesen a súlypont és a legközelebbi pontok. A súlypont referenciapont a szoba geometriai középpontja, amely globális jellemző. A legköze-

lebbi pontok a helyi referenciapontok, ezért a legközelebbi pont a két szoba elhelyezésétől függ. Ez lehet egy szoba másikhoz legközelebbi sarokpontja, vagy két párhuzamos vonalszakasz bármely pontja. A beltéri környezetet ábrázolhatnánk egy gráfon, ahol a csomópontok a helyiségek, és az élek az összeköttetések. A gráfon két csomópont távolsága a legrovidebb úttal mérhető. A két csomópont közötti legrovidebb út különböző algoritmusokkal található meg.

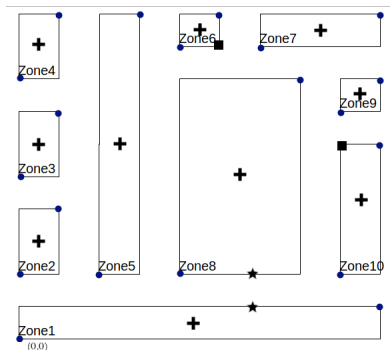
4.2.3. Kísérletek

A javasolt hibaszámítási módszer alkalmazhatósága két kísérleten keresztül kerül bemutatásra. Ezért két kísérleti esetet állítottunk fel különböző referenciapontokkal, azonos távolság- és kapacitásfüggvénnyel. A távolság és a területértékekkel, a tömegvonzás, majd a δ értékek kiszámíthatók. A δ értékek és a Zónák területei alapján hibamátrix készíthető el.

Szimuláció

A tesztkörnyezetet kétdimenziós térben adtuk meg. A környezet 10 zónából áll, és feltételezzük, hogy azok téglalap alakúak, átfedés nélkül, ami az 4.4 ábrán is látható.

A zónákat két átlósan ellentétes pontjuk határozza meg,



4.4. ábra. A Tesz Környezet Elrendezése

és pontokkal jelöltem őket a 4.4 ábrán. A zónák koordinátáit egységekben mérjük. Mindkét kísérletben a zóna területe lett a kapacitásfüggvény. A két zóna távolságát az euklideszi távolságfüggvény alapján számítjuk ki. A referenciapontok kiválasztása megkülönbözteti a két kísérletet, nevezetesen a súlypont, és a legközelebbi pontok kísérlete. Az egyes zónák súlypontját kereszt jelöli az 4.4 ábrán, a Z_6 és Z_{10} pár legközelebbi pontjait négyzetekkel jelöljük, a Z_8 és Z_1 pár esetében csillagokkal vannak jelölve.

Kísérleti eredményeink alapján a bemutatott módszer a zónaméretekkel arányos hibaértékeket eredményez, figyelembe veszi a zónák elrendezését, és két zóna hibaértékei

aszimmetrikusak. Ezért elmondható, hogy a bemutatott osztályozási hiba számítási módszer figyelembe veszi a topológiát az osztályozási hiba kiszámításakor.

Valós életbeli kísérlet

A javasolt tömegvonzás alapú módszer alkalmazhatóságának további teszteléséhez a kísérletet valós életben is elvégezzük. Az IndoorGML (Indoor Geographic Markup Language) [6, 1] az Open Geospatial Consortium (OGC) által meghatározott szabvány, amely egy nyílt formátum a topológia leírására. Ez a dokumentum kétféle formában írja le a helyiségeket háromdimenziós térben, az egyik csúcsokkal való leírás, a másik pedig egy keresésgyorsítást célzó befoglaló doboz. Ezért az IndoorGML lehetőséget nyújt a koordináta-alapú távolságszámítás használatára. Ezenkívül a dokumentum a helyiségek közötti átmeneteket tartalmaz, amelyek leképezhetők gráfra.

Az Informatikai Intézet épülete IndoorGML szabvány segítségével modellezésre került. A koordináták a rendelkezésre álló rács alapján kerültek meghatározásra, így az adathalmaz és az IndoorGML ugyanazt a koordináta-rendszert használja. A háromdimenziós alsó és felső sarok koordinátákat használjuk a távolság és a kapacitás kiszámításához.

A távolság kiszámításához a súlypontok euklideszi távolsága lett választva, és a kapacitás a térfogat lett. A

számított hibaértékek nagyon alacsony átlaggal rendelkeznek, és a szórásnak nagyobbak kell lennie, és a legnagyobb hibaérték alacsonyabb, mint a lehetséges tartomány fele. Ezért a tömegvonzás alapú megközelítés alternatívája lehet a CRISP-megközelítésnek a szimbolikus beltéri pozicionálási módszerek értékelésében.

4.2.4. Összehasonlítás

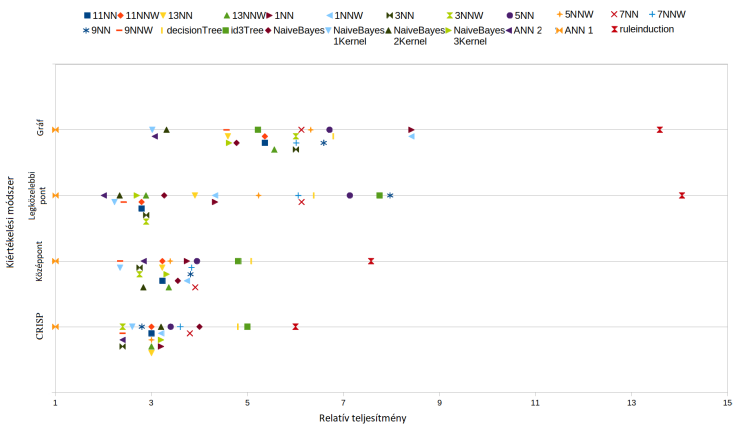
Kísérleteket hajtottunk végre a CRISP és a javasolt topológia alapú osztályozási kiértékelési módszer összehasonlítása céljából. Az összehasonlítást a Miskolc IIS Hibrid adathalmazon végeztük és az épület topológiája IndoorGML szabvánnyal került leírásra. Ugyanazok az osztályozók lettek rangsorolva a CRISP és a javasolt topológia alapú módszer három variánsa alapján, mint amik az első tételben kerültek bemutatásra.

A javasolt topológia alapú osztályozás kiértékelésének három változatát teszteljük különböző távolságfüggvények alkalmazásával. Az első változat a súlypontok közötti euklideszi távolságot alkalmazza. A második változat a legközelebbi pontok euklideszi távolságfüggvényével számított távolságát veszi alapul. A harmadik változat Dijkstra legrövidebb út algoritmusát használja az IndoorGML-ből generált gráfon.

Eredmények

A tesztelt osztályozókat a számított osztályozási hiba alapján rangsoroltuk. Egyező osztályozási hibaértékek esetén az osztályozókat csoportokba rendezzük. A CRISP értékelés legnagyobb csoportját 5 osztályozó alkotja, a gráfon számított távolság esetén 4 osztályozót, míg a súlypontok távolsága és a legközelebbi pontok távolsága egyaránt a 2 osztályozóval rendelkező legnagyobb csoportot eredményezte. Elemeztük az osztályozók rangon alapuló sorrendjét is. Kétféle osztályozó különböztethető meg. A rögzített helyzetű osztályozók alkotják az első típust, a második csoportot a különböző helyzetű osztályozók alkotják. A változatos helyzetű csoportba sorolható osztályozók között vannak olyanok, amelyek jelentős jelentősen eltérő helyen vannak a különböző tesztelt módszerek rangsorában. Az összehasonlításban a kissé eltérő helyzetű osztályozók a leggyakoribbak.

Az osztályozók relatív teljesítménye az értékelési módszer érzékenységének tekinthető. Az egyes értékelési módszereknél az osztályozási hibaértékeket elosztjuk a legjobb értékkel, amely az adott értékelés legalacsonyabb hibaértéke. Az osztályozóknak ez a relatív teljesítménye látható a 4.5 ábrán, ahol a x tengely azt a szorzót mutatja, ahányszor az osztályozónak több hibája van, mint az értékelési módszer legjobb osztályozójának.



4.5. ábra. A Vizsgált Osztályozók Relatív Teljesítménye az Egyes Kiértékelési Módszerek Esetén

A legközelebbi pontok távolságát és a gráf alapján számított távolságát használó módszerek jobban hasonlítanak egymásra, míg a CRISP módszer és a súlypontok közötti távolságon alapuló módszer is nagy hasonlóságot mutatnak. A legközelebbi pontok távolságát alkalmazó módszerrel a legnagyobb a legjobb és a legrosszabb módszer közötti tartomány. Általános eredményeként elmondható, hogy a tömegvonzás alapú megközelítés gyengén teljesítő változata is képes volt kis mértékben felülmúlni a CRISP-megközelítés eredményeit.

4.3. Javított Osztályozás

3. Tézis *Hierarchikus klaszterezéssel kiegészítettem osztályozáson alapuló szimbolikus beltér helyzetmeghatározási módszereket, amelyek így figyelembe veszik az épület topológiáját az osztályozás megbízhatósága alapján.*

A jól ismert osztályozók megbízhatósági értékek alapján adnak eredményt. Vannak olyan esetek, amikor az egyes osztályok megbízhatósági értéke viszonylag kicsi. Ezen osztályozók teljesítményének növelése érdekében bevezethető az osztálykategóriák hierarchikus csoportosítása szimbolikus beltéri helymeghatározás céljából. Topológia leírá-

sával hierarchikus klaszterező algoritmusok alkalmazhatóak a létrehozására.

A szobák az IndoorGML szabvány segítségével kerülnek reprezentálásra. Az IndoorGML dokumentumot át kell alakítani, hogy klaszterezési célokra lehessen használni. A befoglaló doboz alsó és felső sarka, valamint a helyiségek minden csúcsa megtartásra kerül. Az átmenetek nem szükségesek a szobák közötti hierarchia kialakításához, ezért elhagyásra kerülnek. A felesleges jellemzők kiküszöbölése után a kapacitás új funkcióként kerül bevezetésre. Ezenkívül az azonosító és a név jellemzők indexelés céljából összeolvadnak.

A csoportosítás összevonó módon történik két távolságfüggvény felhasználásával, nevezetesen az euklideszi és a tömegvonzás alapú megközelítéssel. A tömegvonzásra épülő megközelítés során bevezetett *delta* virtuális távolságnak tekinthető a szobák között. Az euklideszi távolságot, illetve a térfogatot, mint kapacitásfüggvényt használjuk a tömegvonzás alapú megközelítésnél. Az átlagos, a súlyponti, a teljes, a medián, az egyszeres, a őrségi és a súlyozott összekapcsolási módszereket teszteltük. A súlyponti, a medián és az őrségi összekapcsolási módszerek nem alkalmazhatók tömegvonzás alapú távolsággal, vagy bármilyen másik, az euklideszi távolságtól eltérő távolságfüggvényekkel.

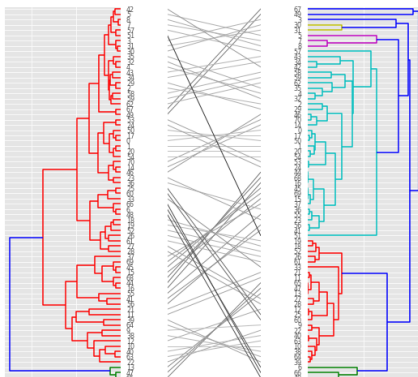
A létrehozott klaszter modelleket a generált dendrogra-

mok alapján összehasonlítjuk. A Tanglegram segítségével össze lehet hasonlítani a fadiagramokat, amelyek összefonódásként mérik a két dendrogram-illesztés minőségét. Az összefonódás értéke az értékelés alapja.

Az azonos távolságfüggvény által létrehozott, de különböző összekapcsolási módszerrel generált fákat összehasonlítjuk egymással, majd optimalizáljuk az összefonódási érték minimalizálása érdekében. Ezek az összefonódási értékek nem feltétlenül szimmetrikusak a megvalósított optimalizáló megközelítés miatt. A kísérleti eredmények azt mutatják, hogy a kapcsolási módszerek többsége hasonló átlagú összefonódási értékeket értek el. Az egyszeres összekapcsolási módszer azonban jelentős különbséget mutat mindkét távolságfüggvényben.

Ezenkívül a két távolságfüggvény megvizsgálható ugyanazon összekapcsolási függvény használatával. A kompatibilitás és az alacsonyabb összefonódási értékek miatt az összehasonlításhoz a súlyozott kapcsolást választottuk. A 4.6 ábra azt mutatja, hogy a két dendrogram nagyon eltérő. Az euklideszi távolság hasonló hierarchikus klaszterezést eredményezhet a legtöbb összekapcsolási módszerrel. Ezek az összekapcsolási módszerek azonban megmagyarázhatónak tűnnek az adott cél szempontjából, a tömegvonzás alapú távolság tükrözheti a topológiát, a szobák elrendezését a dendrogram távolságai szerint.

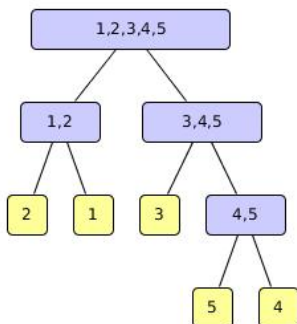
Euklideszi és Tömegvonzás alapú távolság súlyozott összekapcsolási módszerrel és optimalizált összefonódási értékkel: 9.3803



4.6. ábra. Az Euklideszi és a Tömegvonzás alapú Távolság Összefonódási Diagramja Súlyozott Összekapcsolási Módszerrel és 10000 Iterációval

4.3.1. Elgondolás

A szimbolikus pozíciók hierarchikus klaszterezési információinak felhasználásával alacsony megbízhatósági szint esetén javítható a szimbolikus beltéri helymeghatározó algoritmusok pontossága. A javítptt osztályozás koncepciójához osztályozó, küszöbérték és dendrogram szükségesek. A küszöb egy valós érték 0 és 1 között, amely meghatározza, hogy az adott eredményt elfogadjuk-e, vagy a javasolt koncepciót használjuk-e.



4.7. ábra. Az Elgondolás Alap Struktúrája

A hierarchikus klaszterezés által létrehozott fa struktúra a 4.7 ábrán látható. A levélcsomópontok a helyiségek, míg a gyökércsomópont a teljes leírt környezet. A fa

szerkezetét úgy módosítottuk, hogy további információkat tartalmazzon. Minden csomópont tartalmaz mutatókat a szülő és a gyermek csomópontjai számára. A csomópontok adatobjektumot tartalmaznak, amely három információt tartalmaz. Tartalmazza a keresési célú azonosítót, a benne lévő zónák halmazát és e lista méretét.

A továbbfejlesztett fa struktúra alapján a továbbfejlesztési koncepció a következő folyamatot hajtja végre.

1. Az osztályozó elvégzi az osztályozást
2. Ha a megjósolt osztály magabiztossága megegyezik vagy meghaladja a küszöbértéket, akkor a folyamat azzal fejeződik be, hogy eredményül visszaadja az osztályt.
3. A fa levélcsomópontja meghatározásra kerül az azonosító alapján
4. Amíg az aktuális vizsgált csomópont bizalma nem éri el a küszöbértéket, vagy amíg el nem éri a gyökércsomópontot.
 - (a) Megvizsgáljuk az aktuális csomópont szülőjét.
 - (b) A megbízhatóságát a leszármazott levélcsomópontok megbízhatósági értékeinek összegeként számítják ki.

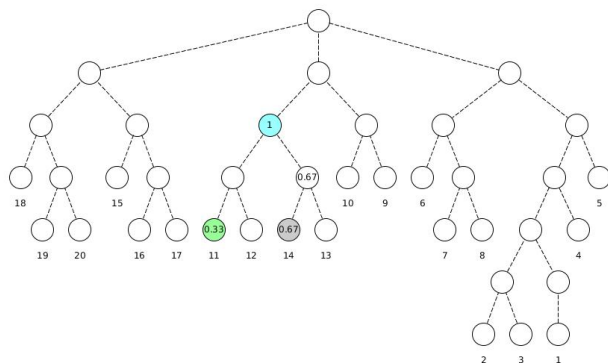
5. A folyamat az utolsó vizsgált csomópont zónáinak visszaküldésével zárul le.

4.3.2. Konceptió igazolása

A javasolt koncepciót egy kísérlet során vizsgáljuk. A kísérlet során a k -NN és a Naive Bayes osztályozókat a rendelkezésre álló funkcionalitáshoz használják, hogy visszaadják az osztály valószínűségeit. Ezek az osztályozók azért lettek kiválasztva, mert könnyen paraméterezhetők. A küszöböt TH -ként van jelölve és $TH \in \{0.6, 0.7, 0.8, 0.9, 1\}$ értékben. Ebben a kísérletben az egyes tesztelt összekapcsolási módszereket minden osztályozóhoz és küszöbértékhez el kell végezni. A távolságfüggvényt úgy választjuk meg, hogy a tömegvonzás alapú megközelítés δ értéke legyen. A kísérlet szűk körben történő vizsgálata során a jobb érthetőség érdekében a környezet a Miskolc Informatika épület IndoorGML-ben definiált második emelete. A Miskolc IIS hibrid adathalmaz azonban ezekből a helyiségekből mindössze 5 szobában végzett méréseket tartalmaz.

Különböző esetek találhatók a tesztben, amelyek bemutatják a koncepció előnyeit. A 9-NN osztályozót súlyozott szavazás nélkül változatát használták az osztály előrejelzésére a mért értékek felhasználásával. A létrehozott dendrogram alapján, súlyozott összekapcsolással és tömegvonzás alapú távolsággal, egy gráf készíthető a 4.8 ábrán

látható módon. Ezen csomópontok közül csak kettőnek van nem nulla valószínűségi értéke, amelyet 11 és 14 jelöl. Az aktuális osztálycsomópontot zöld háttérszín jelöli a grafikonon. Egy hagyományos osztályozó a szürke háttérrel rendelkező csomópontot adja vissza, mivel annak van a legnagyobb valószínűségi értéke.



4.8. ábra. Példa a Javítás Előnyének Bemutatására

A koncepció azonban a megjósolt osztály visszaadása helyett ellenőrzi, hogy a megjósolt osztály valószínűsége eléri-e az adott küszöböt. 0,7 vagy annál magasabb küszöbérték mellett a továbbfejlesztett osztályozó megtalálja az előrejelzett osztályt a grafikonon, és megvizsgálja a szülőjét. Mivel a megjósolt csomópontnak csak egy testvére

van, az is nulla valószínűséggel, a szülő valószínűségi értéke szintén a küszöb alatt van. Emiatt a keresési folyamat egy szinttel feljebb lép a szülő felé. Minden leszármazott levélcsomópont valószínűségének összege 1, amely meghalad bármilyen küszöböt. Így az utolsó vizsgált csomópont, kék háttérrel, a befejező csomópont, amely visszaadja az összes utód levélcsomópontjainak listáját. Az eredeti osztály, a befejező csomópont leszármazottja. Így a továbbfejlesztett koncepció helyesen osztályozta a mérést, csupán 4 szoba felhasználásával.

4.3.3. Eredmények

Az osztályozás vizsgálatához az egyes beállítások három tulajdonságát vizsgálják, nevezetesen a találati arány, a magabiztosság és az absztrakció. A találati arány a helyesen osztályozott esetek aránya és az összes esethez, amely a pontosságot képviseli. Ezért a találati arány valós szám a $[0, 1]$ intervallumban. A célfüggvény a találati arány maximalizálása. A magabiztosság a küszöb és az 1 közötti valós érték, mindkét értéket beleértve, ami az eredmény elfogadott bizalmát képviseli. A célfüggvény a megbízhatósági értékek maximalizálása. A kapott lista méretének minimalizálása érdekében bevezetjük az absztrakció jellemzőt. Ahhoz azonban, hogy összhangban legyen a találati arány és a magabiztossággal célfüggvényeivel, az absztrak-

ció céljának a maximalizálásnak is kell lennie. A helyiségek számának az jellemzőből való kiküszöbölése érdekében az absztrakció szintjét úgy tervezzük, hogy az valós érték a $[0, 1]$ tartományban legyen.

Amikor a pontosság a fő cél, a koncepció az összes szobát vissza tudja adni, aminek eredményeként alacsony absztrakciós szint jön létre. Sőt, ha az absztrakció szintjét a lehető legalacsonyabbra kívánjuk tenni, az osztályozás teljesítménye gyenge lehet. Ezért a küszöb és az összekapcsolás nem csak ezen jellemzők egyikén alapulhat. Az egyensúly megtalálása érdekében egy jósági függvényt vezetünk be ezen funkciók használatával. A bevezetett jósági függvény a 4.4 egyenletben látható, ahol w az adott tulajdonság súlyát jelöli. A cél az jóság értékének maximalizálása.

$$fitness = w_h \cdot talalatiArany + w_c \cdot \overline{magabiztosság} + w_a \cdot \overline{absztrakció} \quad (4.4)$$

A különböző összekapcsolási módszerek észrevehető különbségeket eredményeztek a jósági értékekben. Az eredmények alapján a súlyozott és az egyszeres összekapcsolási módszerek tűnnek jó jelöltnek. A Naive Bayes osztályozó egyik esete sem használta a koncepciót. Ezért a nagyobb körű vizsgálat elfogadható.

4.3.4. Valós életbeli kísérlet

Valós életbeli kísérletet hajtunk végre a javasolt koncepció lehetőségeinek vizsgálatára. Hasonló az előző kísérlethez, de eltér a távolságfüggvény és a környezet tekintetében. A távolságfüggvényt az általánosan használt euklideszi távolságnak és a tömegvonzás alapú megközelítés különbségértékének választjuk. Az eredményeket a találati arány, a magabiztosság és az absztrakció szempontjából is megvizsgáljuk.

Több kritérium alapján meghatározhatjuk a legjobban teljesítő kombinációkat. Például azok az esetek, amelyek a kísérlet során legalább egyszer az egész környezetet adta eredményül, kiszűrésre kerülnek. Egy másik példa: olyan esetek kiküszöbölése, amikor az átlagos halmazméret nem a $(1, 10]$ tartományban van. Ezután az adathalmazt külön vizsgáljuk a távolságfüggvény alapján. Az eredmények alapján a koncepció növelheti a jól ismert osztályozók pontosságát, míg az előre jelzett osztályok száma ésszerű.

Irodalomjegyzék

- [1] Ogc indoorgml-with corrigendum.
<http://docs.opengeospatial.org/is/14-005r4/14-005r4.html>. [Online; accessed 13-October-2017].
- [2] Christian Becker and Frank Dürr. On location models for ubiquitous computing. *Personal and Ubiquitous Computing*, 9(1):20–31, 2005.
- [3] Ulises Carrasco, Pedro Daniel Urbina Coronado, Mahmoud Parto, and Thomas Kurfess. Indoor location service in support of a smart manufacturing facility. *Computers in Industry*, 103:132 – 140, 2018.
- [4] Jim Davis, Thomas Edgar, James Porter, John Bernaden, and Michael Sarli. Smart manufacturing, manufacturing intelligence and demand-dynamic perfor-

- mance. *Computers & Chemical Engineering*, 47:145 – 156, 2012. FOCAPO 2012.
- [5] Heiner Lasi, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann. Industry 4.0. *Business & information systems engineering*, 6(4):239–242, 2014.
- [6] J Lee, Ki-J Li, S Zlatanova, TH Kolbe, C Nagel, and T Becker. Ogc® indoorgml. *Open Geospatial Consortium standard*, 2014.
- [7] M. Lichman. UCI machine learning repository, 2013.
- [8] Shaoping Lu, Chen Xu, Ray Y. Zhong, and Lihui Wang. A rfid-enabled positioning system in automated guided vehicle for smart factories. *Journal of Manufacturing Systems*, 44:179 – 190, 2017.
- [9] Isaac Newton. *The Principia: mathematical principles of natural philosophy*. Univ of California Press, 1999.
- [10] Joaquin Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Joan P Avariento, Tomás J Arnau, Mauri Benedito-Bordonau, and Joaquin Huerta. Uji-indoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization

- problems. In *Proceedings of the fifth conference on indoor positioning and indoor navigation*, 2014.
- [11] Joaquin Torres-Sospedra, David Rambla, Raul Montoliu, Oscar Belmonte, and Joaquin Huerta. Ujiindoorloc-mag: A new database for magnetic field-based localization problems. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, pages 1–10. IEEE, 2015.
- [12] Zsolt Toth. Ilona: indoor localization and navigation system. *Journal of Location Based Services*, 10(4):285–302, 2016.



Nyilvántartási szám: DEENK/330/2020.PL
Tárgy: PhD Publikációs Lista

Jelölt: Kunné Tamás Judit
Doktori Iskola: Informatikai Tudományok Doktori Iskola
MTMT azonosító: 10060570

A PhD értekezés alapjául szolgáló közlemények

Idegen nyelvű tudományos közlemények hazai folyóiratban (1)

1. Tóth, Z., Magnucz, P., Németh, R., **Tamás, J.**: Data Model for Hybrid Indoor Positioning Systems.
Prod. Syst. Inf. Eng. 7, 67-80, 2015. ISSN: 1785-1270.

Idegen nyelvű tudományos közlemények külföldi folyóiratban (2)

2. **Tamás, J.**, Tóth, Z.: Topology-Based Evaluation for Symbolic Indoor Positioning Algorithms.
IEEE Trans. Ind. Appl. 55 (6), 6324-6331, 2019. ISSN: 0093-9994.
DOI: <http://dx.doi.org/10.1109/TIA.2019.2928489>
IF: 3.488
3. **Tamás, J.**, Tóth, Z.: Classification-based symbolic indoor positioning over the Miskolc IIS Data-set.
J. Locat. Based Serv. 12 (1), 2-18, 2018. ISSN: 1748-9725.
DOI: <http://dx.doi.org/10.1080/17489725.2018.1455992>

Magyar nyelvű konferencia közlemények (1)

4. **Tamás, J.**, Tóth, Z.: Osztályozáson alapuló pozicionálási módszerek vizsgálata a Miskolci Informatikai Épület Hibrid Adathalmazára alapján.
In: Műszaki tudomány az Észak-Kelet Magyarországi régióban 2016 : konferencia előadásai.
Szerk.: Bodzás Sándor, Debreceni Akadémiai Bizottság Műszaki Szakbizottsága, Debrecen, 664-668, 2016. ISBN: 9789637064333

Idegen nyelvű konferencia közlemények (7)

5. **Tamás, J.**, Tóth, Z.: Classification Refinement With Category Hierarchy.
In: Proceedings of the 11th International Conference on Applied Informatics (ICAI 2020). Ed.: by Gergely Kovácsnai, István Fázekas, Tibor Tóth, CEUR, Eger, 358-369, 2020, (CEUR Workshop Proceedings, ISSN 1613-0073 ; 2650.)
6. **Tamás, J.**: Hierarchical Clustering based on IndoorGML Document.
In: 2019 IEEE 15th International Scientific Conference on Informatics : Informatics 2019 Proceedings. Eds.: William Steingartner, Stefan Korecko, Anikó Szakál, IEEE, Piscataway, 411-416, 2019. ISBN: 9781728131788





7. Ilku, K., **Tamás, J.**: IndoorGML modeling: a case study.
In: Proceedings of the 2018 19th International Carpathian Control Conference (ICCC). Eds.:
Dániel Drótos, József Vásárhelyi, László Czap, Ivo Petrás, IEEE, Danvers, 633-638, 2018.
ISBN: 9781538647622
8. **Tamás, J.**, Tóth, Z.: Limitation of CRISP accuracy for evaluation of room-level indoor positioning
methods.
In: Proceedings of 2018 IEEE International Conference on Future IoT Technologies (Future
IoT), IEEE, Washington, 1-6, 2018. ISBN: 9781538612088
9. Ilku, K., **Tamás, J.**: Topology-based Classification Error Calculation based on IndoorGML
Document.
In: The 11th Conference of PhD Students in Computer Science : volume of short papers :
CS2, University of Szeged, Szeged, 101-105, 2018.
10. **Tamás, J.**, Tóth, Z.: Topology-based classification error calculation for symbolic indoor
positioning.
In: Proceedings of the 2018 19th International Carpathian Control Conference (ICCC). Eds.:
Dániel Drótos, József Vásárhelyi, László Czap, Ivo Petrás, IEEE, Danvers, 643-648, 2018.
ISBN: 9781538647622
11. Tóth, Z., **Tamás, J.**: Miskolc IIS hybrid IPS: dataset for hybrid indoor positioning.
In: Proceedings of the 26th International Conference Radioelektronika 2016, IEEE,
Piscataway, 408-412, 2016. ISBN: 9781509016747

A közlő folyóiratok összesített impakt faktora: 3,488

**A közlő folyóiratok összesített impakt faktora (az értekezés alapjául szolgáló közleményekre):
3,488**

A DEENK a Jelölt által az iDEa Tudóstérbe feltöltött adatok bibliográfiai és tudománymetriai
ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján
elvégezte.

Debrecen, 2020.11.02.

