AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY

Associative Graph Database System

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

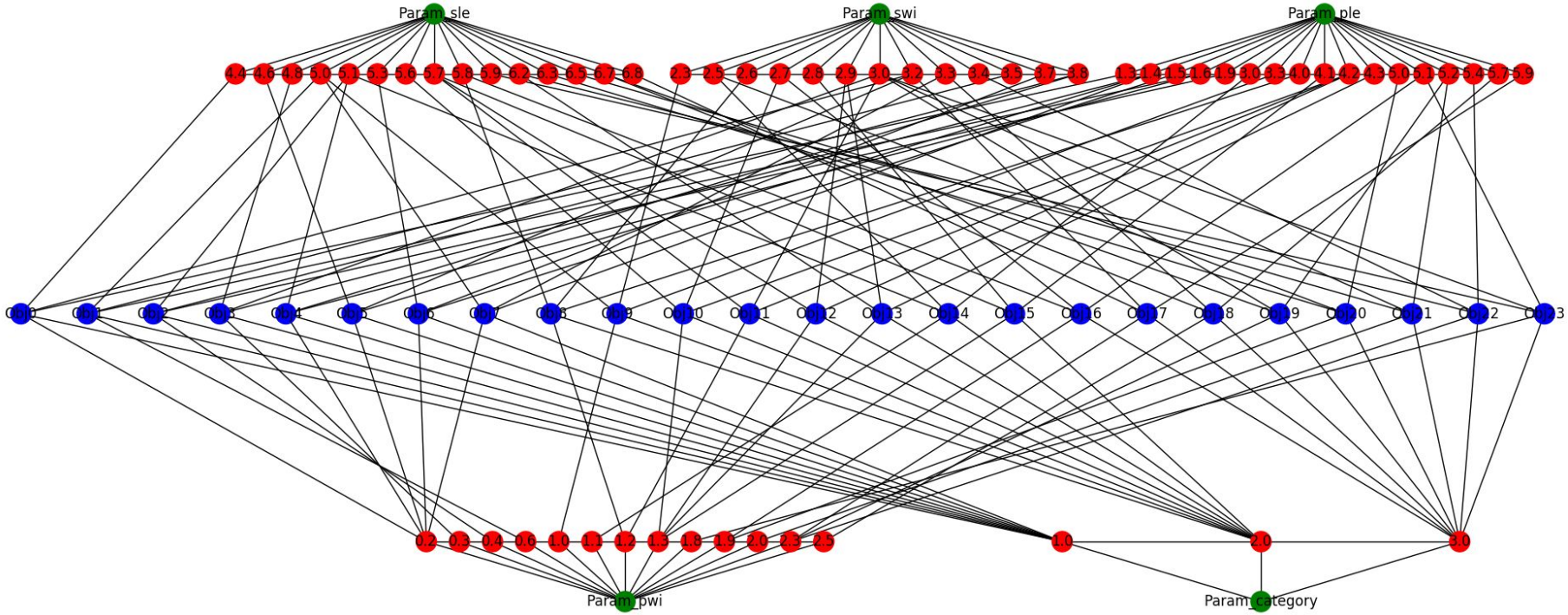
Krzysztof Ćwieka

Kraków, 18.09.2019

Construction of an associative graph database system

1. Given a database in form of a table with records and their parameters
2. For each of the parameters the value duplicates are removed.
3. Each piece of the remaining table is assigned its own node.
4. Edges are created between:
   a. object nodes and their parameter values
   b. parameter values and nodes representing names of parameters.
   c. parameter values that are neighbours in value

Example of a database



Weights of edges

Edges are assigned weights according to the following rules:

- between objects and values - the weight is defined as inverse of the number of object nodes connected to a given value node
- between neighbouring values - the weight is defined as absolute difference between these two values divided by the range of a given parameter

Similarity

Given a set of values to calculate similarity to, a similarity is calculated in a following manner:

Each parameter of the record in question is assigned similarity of 1. Then each neighbouring value node is assigned similarity of known similarity times weight, until all parameters have their similarities assigned. Then each object has its similarity calculated as a sum of similarities of its values divided by the number of parameters.

Classification algorithms

Classification is a task, where given a dataset with some parameters and each record assigned to some class, a record with no known assignment can be classified as belonging to some class using an algorithm.

In this project three algorithms were used:

1. Mean similarity
2. k Nearest Neighbours
3. Fast k Nearest Neighbours

Mean similarity

In mean similarity algorithm a similarity is calculated for each and every record in the database, then records are sorted according to their class and for each class a mean similarity is calculated by summing similarity rates of all records in a given class by the number of them. The class with the highest score is assigned to the record in question.

k Nearest Neighbours

In k nearest neighbours algorithm similarities are calculated for each and every record in the database, then they are sorted according to this similarities and k highest scores are selected. The class assigned to the record in question is the one that is most frequent in k highest scores.

kNN Fast

Fast k nearest neighbors is similar to kNN algorithm except only a limited number of closest records is taken into account in order to make the algorithm faster. Some number is chosen to limit how many neighbours of values of a record in question are having their similarities calculated. Then only objects connected to those values have their similarities calculated.

Accuracy comparison of the algorithms

Classification accuracy results (in percent) with k=5 (for kNN and kNNFast) on test datasets:

Dataset\Algorithm	Mean similarity	kNN	kNNFast
Iris	58	100	100
Wine	45	62	70
Wisconsin Cancer	25	80	96
Glass Identification	15	30	55

Time comparison of the algorithms

Results on Iris dataset, other sets produce similar relations between execution times.

Algorithm	Time (s)
Mean similarity	0.034
kNN	0.105
kNNFast	0.085

Additional implemented functions

- calculating mean value of a given parameter
- calculating median value of a given parameter
- calculating similarity to a given object from database or a new user-provided parameters

Conclusions

Although mean similarity method is the fastest it provides the worst classification accuracy. K nearest neighbours and its fast version provide similar accuracy results, however kNNFast is both faster of the two and more accurate.

THE END