Bachelor thesis

Artificial Intelligence

Radboud University

Using semantic and graph similarity methods to predict heart or kidney abnormalities

Author:

Misja Zippelius

S1019965

First supervisor:

A.J.M. Dingemans

Department of Human Genetics, Donders Institute for Brain, Cognition and Behavior

a.dingemans@radboudumc.nl

Second supervisor:

Dr M. Hinne

Department of Artificial Intelligence, Donders Institute for Brain, Cognition and Behavior

m.hinne@donders.ru.nl

IN·DEI·NOMINE·FELICITER

January 30, 2022

Abstract

There are many symptoms, which cannot be visually spotted. Imaging is needed to spot these symptoms, which is time intensive. Two examples of these symptoms are heart abnormalities and kidney abnormalities. These can be predicted with the use of graph and semantic similarity methods. Patients are represented as HPO graphs with symptoms being HPO nodes in the graph. Graph and semantic similarity can be used to compare patients with each other in order to make a classification based on the most similar patients. Resnik and Lin, two semantic similarity methods, have been compared to the maximum common sub-graph method, which is a graph similarity method. Overall the maximum common sub-graph method yields the best performance and has an AUC of 0.89. This can already be used with this data to determine the priority of the order in which patients undergo imaging.

Contents

Chapter 1

Introduction

In the medical world a doctor helps many patients every year. Take for example a geneticist; for each patient he/she wants to know whether the patient has a syndrome and which symptoms the patient has. Some symptoms, like facial abnormalities, can be visually observed by a geneticist, others like heart or kidney abnormalities cannot. To discover these symptoms, imaging has to be used, which can be time intensive if it needs to be done for all patients. Especially if the hospital in question is located in a developing country, where imaging machines are scarce[1]. In this case, the hospital needs to make a well thought-out decision as to which patients should undergo imaging. Additional information about the patient can be used to save time and make such a well thought-out decision. This additional information can be the symptoms a patient has, for instance ventricular septal defect. Especially in genetic syndromes, where a patient can have a wide range of symptoms, it is conceivable that the presence of some symptoms can be predictive of the presence of others.

The difficulty however when using symptoms, is inter-observer bias. This is the reason the Human Phenotype Ontology (HPO) terms were developed. Ontology means a vocabulary that is standardized and in this case it contains many terms with phenotypic information of human genes. These terms which contain phenotypic abnormalities are structured in graphs, which are directed and in which there are no cycles possible[2]. The graphs start out with abstract terms at the top and get more specific the deeper you get into the graph. So a very specific abnormality such as ventricular septal defect will be very deep in the graph whereas a more general abnormality like heart abnormality will be less deep in the graph.

To compare patients based on their HPO terms, graph similarity could be used. Graph similarity is a concept, which is widely used in many domains especially since the rise of the world wide web, which is a large graph that consists of billions of web pages[3]. Graph similarity allows us to compare graphs and rank them on the basis of similarity. There are many different graph similarity measures, each with their own advantages. Graphs are found in many daily tasks, such as task scheduling and social networking[4] [5].

Graph similarity can also be useful with HPO graphs, specifically in the genetics domain. Doctors can spare a lot of time and resources if symptoms, which are not easily detectable by the geneticist could be predicted. This can be done if the patients are classified into the positive or negative group for the specific symptom of interest. HPO graphs can be used for this classification task along with graph similarity. There are many different graph similarity measures, but there exist semantic similarity measures as well, which are specific for an ontology. Semantic similarities measure how similar certain concepts are in one specific ontology[6]. So the research question answered in this thesis is:

Which graph or semantic similarity measure works best when predicting whether patients have heart or kidney abnormalities based on their phenotype?

On the basis of these HPO graphs, we can compare the patient of interest to other patients from the dataset. Intuitively, graph and semantic similarity can be used to compare patients to each other and make a list of most similar patients. In order to construct a similarity matrix, two semantic similarity measures and one graph similarity measure are used. Afterwards a classification rule can be defined, which will determine the likelihood of a positive class for each patient. This likelihood will be transformed into a probability for each patient, which stands for the probability that each patient belongs to the class of interest. This class of interest is in our case the symptom we want to predict. The model can be tweaked on the basis of the doctors preferences, which determines the classification threshold. False negatives are a problem in the medical world, because they it means that patients who have the disease will not get recognized by the classifier. In order to minimize the false negatives to zero, we can adapt the threshold to a certain point. This will lead to additional false positives, meaning that patients will be classified as if they have the disease when they do not. However, this solution is still able to classify all the positive patients correctly and some of the negative patients as well, which will lead to a decrease in the amount of times imaging needs to be used.

The research uses HPO data to asses similarity between patients as is also done by Rojano et al.[7]. It also uses graph similarity for a classification task, which is also done by Diykh et al.[8]. The research combines and compares it with semantic similarity methods, which are also used for classification by Rojano et al.[7]. This research tries to predict heart and kidney abnormalities. This has also been done earlier by Mohan et al.[9] and Vijayarani et al.[10], but contrary to these studies, this research uses HPO data.

Chapter 2

Preliminaries

2.1 Graphs

Graphs are data structures, which are built from nodes and edges. A node is point in a graph, which can be connected with other nodes through edges. An edge is a connection between two nodes. Graphs can represent many things such as a social network with people representing nodes and edges the connections between people. There are two types of graphs: directed and undirected graphs. These concepts tell something about the direction of traversal in a graph. An undirected graph consists of edges which are bidirectional, meaning that it is possible to traverse the edge in both directions whereas it is only possible to traverse the edge in one direction with directed graphs. Nodes can also be called vertices and will be denoted as V. Edges are denoted as a connection between two vertices. For example, an edge that connects V1 and V2 is denoted as E(V1,V2). A sub-graph is a graph G2, which is part of a larger whole graph G. This means that the set of vertices in the sub-graph is a subset of the set of vertices in the whole graph and that the set of edges in the sub-graph is a subset of the set of edges in the whole graph[11] [12].

2.2 Graph Isomorphism

If two graphs are isomorphic it means that both graphs have the same amount of nodes, edges and that those nodes are connected in the same way in both graphs[13].

2.3 Graph Similarity - Maximum Common sub-graph

Graphs are useful in many domains. Graphs can also be compared to each other to assess how similar they are. Graph similarity is an ambiguous concept and it can be interpreted in many ways, thus there are numerous notions of comparison. In this research the maximum common sub-graph method is used, which is an example of isomorphism with sub-graphs. If we have two graphs, common sub-graphs are the sub-graphs that the two graphs have in common and thus are isomorphic. Out of all the common sub-graphs the two graphs have, the maximum common sub-graph (MCS) is the sub-graph with the highest amount of edges. In this research I will only look at MCS's which are connected, meaning there is a path between every pair of distinct nodes[14] [15].

2.4 Resnik And Lin Similarity

The Resnik and Lin similarity are both semantic similarity measures. A semantic similarity measure is a metric for the similarity of terms in an ontology. It calculates a score based on the resemblance between the meaning of both terms[16]. The Resnik similarity measure is based on the information content (IC) of the terms. Two terms have a higher similarity if they share more information. The information content between two terms is defined as: -log p(c). p(c) is the probability of encountering an instance of concept c. Within a set of terms an abstract concept has a high probability, but a low information content. On the other hand a specific concept has a low probability, but a high information content. The similarity between two terms is defined as:

$$Sim(c1, c2) = max_{c \in S(c1,c2)}[-log\ p(c)]$$

S(c1,c2) is the set with the common ancestors of c1 and c2. The similarity of two terms is equal to the concept in S with the highest information content, also called the most informative subsumer[17].

The Lin similarity in is defined as:

$$Sim(x1, x2) = \frac{2 \cdot LogP(C_0)}{LogP(C_1)+LogP(C_2)}$$

The C0 stands for the most specific common ancestor for x1 and x2. C1 and C2 are respectively class 1 and 2. Calculating the Resnik for x1 and x2 will lead to the same result, so the above formula is equal to:

$$\frac{2 \cdot ResnikSimilarity(c1,c2)}{IC(c1)+IC(c2)}$$

[18]

Chapter 3

Related Work

There has not been many research on the classification of patients with HPO data. However, there are many studies on the prediction of heart and kidney diseases. These studies do not use HPO data, but other healthcare data. Take for example this paper by Mohan et al., which uses machine learning techniques on a data set with many attributes to predict whether someone has a high chance to develop a heart disease. The study uses a hybrid random forest with a linear model to make these predictions. It already achieves an accuracy of 88.7 percent[9]. A similar study by Vijayarani et al. exists in the form of Kidney disease prediction. Here Support Vector Machines and Artificial Neural Networks are used to classify patients into five categories. These categories consist of four kidney diseases and one healthy class. This research also uses healthcare data collected from medical labs, hospitals and centres. The data consists of attributes, which can influence renal diseases[10].

Graph and semantic similarity can also be used for classification. Research by Diykh et al. shows that graph properties and structure can be used for classification. In the study, graph properties, such as the degree distribution and Jaccard Similarity Coefficient. Based on these features, graphs are clustered on their similarity in order to classify sleep EEG signals[8]. Semantic similarity measures can also be used for a classification task. This study by Rojano et al., which uses the Lin method to calculate the similarity between patients and cluster them into cohorts. The study uses HPO data from patients to cluster the patients into genetic disease cohorts. It uses the Lin method to asses the similarity between the patients, such that they can be clustered on the basis of phenotypic similarity[7].

This research expands upon above mentioned studies, by using Human Phenotype ontology data for the prediction of heart and kidney abnormalities. This adds the other symptoms, which can be predictive of heart and kidney abnormalities to the used data. This research also combines the usage of graph and semantic similarity methods for a classification task. These methods are also compared to each other. All in all, this research combines methods and data, which have not been combined before to attempt to solve the problem described in the introduction.

Chapter 4

Research

4.1 Hypothesis

The three similarity methods used are: maximum common sub-graph(MCS), Resnik and Lin. It is suspected that both the Resnik and Lin similarity methods will perform better than the MCS method. Considering they are both semantic similarity methods, which are specific for an ontology, whereas the MCS similarity method is intended for graphs in general.

4.2 Research Method

To find out the influence of the similarity methods on the prediction of heart or kidney abnormalities, the methods were implemented in python and were used to create similarity matrices. Both the Resnik and Lin similarity come from the Phenopy library, which lets us calculate the similarity between two lists of HPO terms[19]. The maximum common sub-graph method was implemented with the assistance of methods from NetworkX, which is a Python library for graphs[20].

The dataset contains 553 patients. Three columns of this dataframe are of importance in this research. The first column 'hpo_all_with_parents' contains lists of all patients and those lists contain all HPO terms of that patient. The second column 'hpo_all' is build up in the same way, but only contains the most specific child nodes. This means that the list only contains nodes that have no descendants. This column will be used to calculate the Resnik and Lin similarity as this is much more computationally efficient than using all HPO terms and produces almost the same results. The third column of importance is 'graphs', which contains the graphs of all patients with all HPO terms in NetworkX format.

First three similarity matrices are calculated: one for each similarity method. The matrices for Resnik and Lin are calculated in almost the same way and the procedure consists of a few steps. At first the program will loop through all the patients in the dataset and for each patient, it is checked whether the patient has heart or kidney abnormalities. This can be done by checking whether the patient in hpo_all_with_parents contains the most abstract HPO term for heart or kidney abnormalities. The terms of interest are in this case HP:0000077, which means 'Abnormality of the kidney' and HP:0001627, which means 'Abnormal heart morphology'. Moreover all the specific descendants of these terms are also of interest as they are more specific instances of kidney and heart abnormalities[21].

Secondly, if a patient has either of these abnormalities or a specific instance of them they will be removed from the data. This procedure simulates a 'patient' where these symptoms have not been discovered. Then we will use Leave-One-Out Cross-Validation; each patient in hpo_all will be compared one by one to all remaining 552 patients with the Resnik and Lin similarity. This gives us a a similarity matrix consisting of 553 entries. This includes a comparison of the patient of interest with itself, which will not be used for the prediction. Ultimately we will end up with a 553 by 553 similarity matrix.

The similarity matrix for the MCS method is computed in a similar way, but uses the 'graphs' column. When calculating the similarity with this method there needs to be a correction for the size of the graphs, because some patients have far larger graphs than others. This is done by dividing the size of the maximum common sub-graph by the size of the patient with which we are comparing with. The size of a graph is here defined by the total amount of nodes and edges. Leave-One-Out Cross-Validation is used as well and the result is a 553 by 553 similarity matrix.

After calculating the similarity matrices, a classification rule needs to be defined. This rule will be used to calculate a score for each patient. Based on these scores the competence of the classifier can be evaluated. A few different classification rules for each similarity method have been compared. For the classification rule we will look at the most similar patients for each patient. Here the variable of interest is the range of most similar patients used to calculate the score. The score is calculated by summing the amount of patients with a heart or kidney abnormality in the top range of most similar patients. This depends on which abnormality we want to predict. For example: The variable is set to 10 and we want to predict heart abnormalities; the program will look at the top 10 most similar patients for each patient and calculate the amount of patients with heart abnormalities in the top 10. In this research, the top 1, 5, 10, 25, 50 and 100 most similar patients have been tested to determine what works best with every similarity method.

The combination of semantic similarity methods and graph similarity methods has also been tested. This can be done by taking the average of the two similarity matrices. Combining all the methods with each other leads to three additional similarity matrices, which are the result of the three different combinations: Resnik and Lin, Resnik and MCS, Lin and MCS. When combining a semantic similarity measure with a graph based similarity measure, the matrix of the semantic measure needs to be scaled from 0 to 1, as that is the range of the values in the MCS matrix. No scaling is needed for the combination of Resnik and Lin.

After the initial steps, it is possible to evaluate the classifier and plot a ROC curve for the classifier. The ROC curve is a graph of the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds.

$TPR = \frac{True\ positives}{All\ positives}$

$FPR = \frac{False\ positives}{All\ negatives}$

If the classifier is random and the probability for the prediction of each class is 0.5, the ROC curve is a straight line from the bottom left to the top right. It is also possible to calculate the area under the ROC curve (AUC) of the classifier. The AUC for the ROC curve tells us how well the classifier performs. An AUC of 1 would be perfect, whereas an AUC of 0.5 means that the classifier is random. Furthermore, the AUC is sensitive to class imbalance, which is desired in this case, because the classes of patients with an abnormality are much smaller[22]. A different measure used for the evaluation of the classifiers is the Brier score, which is the mean squared error of the predicted probability and measures the accuracy of the classifier. A Brier score of 0.25 means a random classifier and a score of 0 would be perfect, since there would be no error in prediction[23]. I will use the Brier and AUC score to asses the performance of the models. For the ROC curve, AUC score and the brier score, the Sklearn library is used [24].

After evaluating the model it can be used. From the ROC curve it is possible to select a threshold according to one's preferences. If the doctor for example wants to make predictions without any false negatives, he can select the point in the ROC graph where the TPR is 1 and use that threshold. There can also be other scenarios. Take for example a hospital, located in a developing country, where they only have one imaging machine, which urges an efficient use of that machine. In this case the threshold most close to the left upper corner can be used, as this is the perfect situation where the TPR is 1 and the FPR 0.

4.3 Results

4.3.1 Data Exploration

The Dataset consists out of 553 patients. From those 553 patients 113 have a heart abnormality and 57 have a kidney abnormality. Figure 4.1 and 4.2 show the ratio of the patients with abnormalities to the total amount of patients.

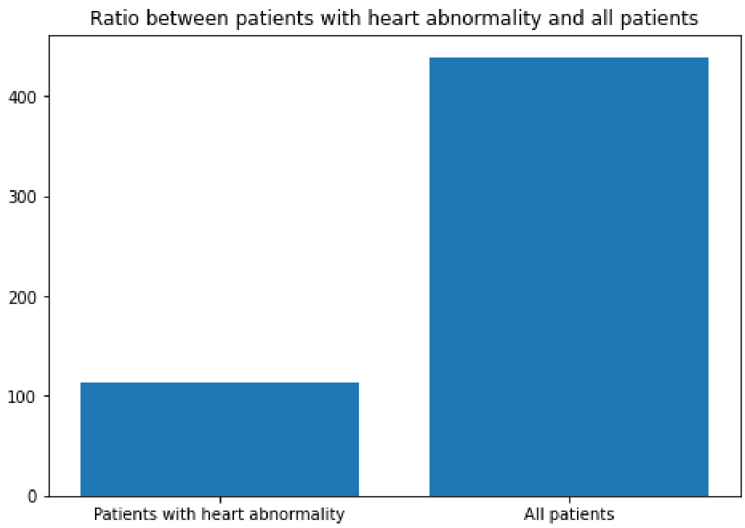

Figure 4.1: A visualization of ratio between the amount of patients that have a heart abnormality and all the patients

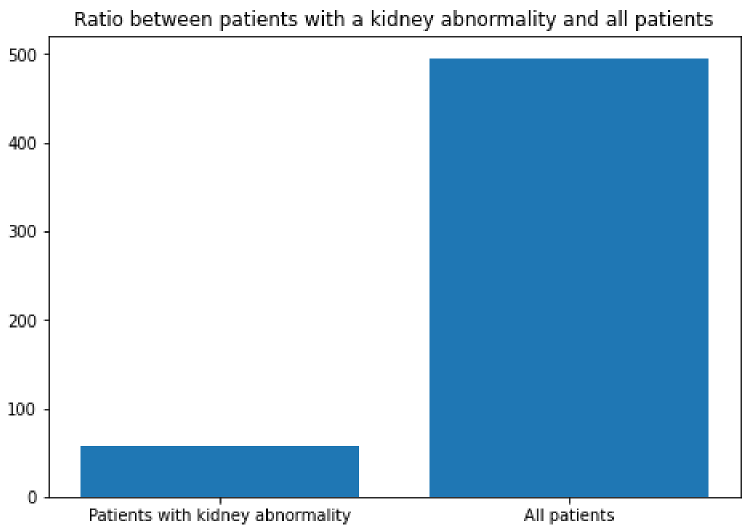

Figure 4.2: A visualization of ratio between the amount of patients that have a kidney abnormality and all the patients

4.3.2 Kidney Abnormality

In this section the results are presented of the case which concerns the prediction of kidney abnormalities. The first method here will be the Resnik similarity. In Figure 4.3, the different amount of top most similar patients are displayed with their AUC scores. The last three scores are close to each other. When examining them closely, considering the top 50 most similar patients yields the best result and leads to a rounded AUC score of 0.76. In figure 4.4, the ROC curve of the Resnik with top 50 most similar patients is plotted. The graphs shows that a TPR of 1 is only achievable with a very high FPR of almost 1, so this scenario is not usable. It is however possible to achieve a TPR of 0.8 and a FPR of 0.3.
This classifier leads to a Brier score of 0.085.

The second method is the Lin similarity, which results are similar to that of the Resnik score. In Figure 4.3, the different amount of top most similar patients has been displayed with the according AUC scores. The last three scores are close to each other. When examining them closely, considering the top 25 most similar patients yields the best result and leads to a rounded AUC score of 0.73. In figure 4.4, the ROC curve of the Lin with top 25 most similar patients is plotted. The ROC curve again shows that a TPR of 1 is only achievable with a very high FPR of almost 1, so this scenario is not usable. It is however possible to achieve a TPR of 0.8 and a FPR of 0.3.
This classifier leads to a Brier score of 0.087, which is a little higher than the Brier score of the Resnik method.

The last method is the MCS method. In Figure 4.3, the different amount of top most similar patients has been displayed with the according AUC scores. In the case of the MCS method, there is one case clearly better; Considering the top 100 most similar patients yields the best result and leads to a rounded AUC score of 0.89. In figure 4.4, the ROC curve of the MCS with top 100 most similar patients is plotted. The ROC curve shows that in this case a TPR of 1 is achievable with a FPR of 0.5, which can be usable in some scenarios. It is also possible to achieve a TPR of 0.82 and a FPR of 0.2.
This classifier leads to a Brier score of 0.086 , which is a little higher than the Brier Score of the Resnik method.

AUC Scores: Kidney Abnormality

	1	5	10	25	50	100
Resnik	0.5	0.62	0.70	0.75	**0.76**	0.74
Lin	0.5	0.64	0.70	**0.73**	**0.73**	0.72
MCS	0.5	0.67	0.78	0.83	0.84	**0.89**

Figure 4.3: AUC scores when predicting whether patients have a kidney abnormality. Three different methods and a variable amount of most similar patients are used.

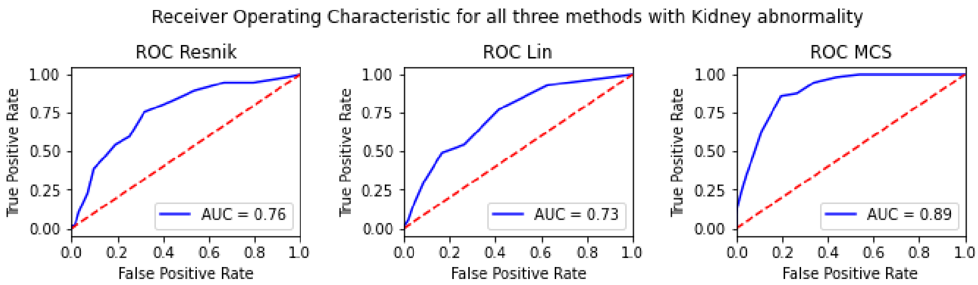

Figure 4.4: ROC curves of three different methods

	Brier Score
Resnik	0.085
Lin	0.087
MCS	0.086

Figure 4.5: Brier Scores of the kidney predictions made with the three different methods

4.3.3 Heart Abnormality

In this section the results are presented of the case which concerns the prediction of heart abnormalities. The first method here will be the Resnik similarity. In Figure 4.6, the different amount of top most similar patients has been displayed with the according AUC scores. The last three scores are close to each other. When examining them closely, considering the top 100 most similar patients yields the best result and leads to a rounded AUC score of 0.72. In figure 4.7, the ROC curve of the Resnik with top 100 most similar patients is plotted. The graphs shows that a TPR of 1 is only achievable with a very high FPR of almost 1, so this scenario is not usable. This classifier leads to a Brier score of approximately 0.149.

The second method is the Lin similarity, which results are similar to that of the Resnik score. In Figure 4.6, the different amount of top most similar patients has been displayed with the according AUC scores. The last three scores are close to each other. When examining them closely, considering the top 50 most similar patients yields the best result and leads to a rounded AUC score of 0.71. In figure 4.7, the ROC curve of the Lin with top 50 most similar patients is plotted. The ROC curve again shows that a TPR of 1 is only achievable with a very high FPR of almost 1, so this scenario is not usable.

This classifier leads to a Brier score of approximately 0.149, which is the same as the Resnik method.

The last method is the MCS method. In Figure 4.6, the different amount of top most similar patients has been displayed with their AUC scores. The last two scores are close to each other. When examining them closely, considering the top 50 most similar patients yields the best result and leads to a rounded AUC score of 0.86. In figure 4.7, the ROC curve of the MCS with top 50 most similar patients is plotted. The ROC curve shows that in this case a TPR of 1 is achievable with a FPR of 0.65, which can be usable in some scenarios. It is also possible to achieve a TPR of 0.73 and a FPR of 0.18.

This classifier leads to a Brier score of approximately 0.152 , which is a little higher than the Brier Score of the Resnik and Lin method.

AUC Scores: Heart Abnormality

	1	5	10	25	50	100
Resnik	0.5	0.66	0.68	0.71	0.71	**0.72**
Lin	0.5	0.65	0.68	0.70	**0.71**	**0.71**
MCS	0.5	0.72	0.75	0.81	**0.86**	**0.86**

Figure 4.6: AUC scores when predicting whether patients have a heart abnormality. Three different methods and a variable amount of most similar patients are used.

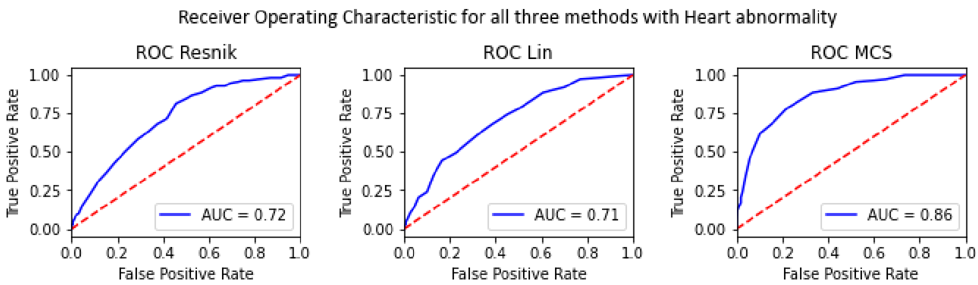

Figure 4.7: ROC curves of three different methods

	Brier Score
Resnik	0.149
Lin	0.149
MCS	0.152

Figure 4.8: Brier Scores of the heart predictions made with the three different methods

Chapter 5

Discussion

The results of this study indicate that overall the maximum common subgraph method works the best when predicting whether patients have heart or kidney abnormalities based on their phenotype. This is contrary to the hypothesis, which suggests that the Resnik and Lin semantic similarity methods would work the best, since they are specific for an ontology.

The data show us that predicting kidney abnormalities works better than predicting heart abnormalities, because The AUC scores of the classifiers are all higher when predicting kidney abnormalities than when predicting heart abnormalities. The analysis also identifies that the maximum common subgraph method works the best when predicting either of these abnormalities, since the classifier has the highest AUC of 0.86 and 0.89, when respectively predicting heart and kidney abnormalities. It also has a low Brier score, meaning that the prediction error is small. This classifier is not perfect, but it could have various uses in society. It could be used to determine the priority of patients, which are on the waiting list for imaging. This can especially be useful in developing countries, where imaging machines are limited or in a crisis situation.

A patient visits the hospital and the doctor wants to know whether this patient has a heart or kidney abnormality. At first focused on the kidney abnormality, the Doctor will enter the patient's known symptoms in the computer as HPO terms. From this, a graph can be easily constructed. From the results of the study it can be inferred that, using the MCS similarity method on the HPO graph will lead to the best performance. This can be derived from it's AUC score, which at 0.89 is the highest of the three classifiers. Using this classifier, the doctor is able to fine-tune the threshold of the classifier to his/her own liking's. The doctor is able to achieve a TPR of 1 with a FPR of 0.5 if he/she wants to prevent any false negatives. In this case, the probability of falsely classifying this patient as positive is 0.3 and the probability of correctly classifying the patient if he is positive, is 1. If the doctor was operating in a developing country in a busy hospital with only a limited amount of imaging machines, he/she might choose to achieve the lowest possible FPR with the highest possible TPR. This point lays closest to the top left corner in the ROC curve. It will have a TPR of 0.82 and a FPR of 0.2. Both the FPR and TPR are lower this time. The probability that this patient is now correctly classified, when he is in fact positive, is 0.82. So 0.18 percent of all patients, who are in fact positive, will be falsely classified as negative. On the other hand, the probability that this patient will be falsely classified as positive, when he is in fact negative is now 0.2. So 20 percent of all negative patients will be classified falsely as positive. If these results are directly related to the data, 10 of the 57 patients with a kidney abnormality will be classified falsely as negative. Out of all the 496 negative patients approximately 100 will be classified falsely as positive.

The brier score of the MCS method compared to the other methods unfortunately does not say much; They lay close to each other. As stated in the results, combining two similarity methods does not lead to a better result.

If the doctor focuses on the heart abnormality, the best classifier is still the one, which uses the maximum common sub-graph similarity method. This classifier will lead to an AUC of 0.86. The doctor is again able to fine-tune the threshold of classification to his/her own requirements. If no false negatives are desired, a TPR of 1 is achieved with a FPR of 0.65. The point, which lays closest to the top-left corner and which represents the most well-rounded combination, has a TPR of 0.73 and a FPR of 0.18. If these results are again directly related to the data, approximately 31 of the 113 patients with a heart abnormality will be classified falsely as negative. Out of all the 440 negative patients approximately 80 will be falsely classified as positive. Once more, the Brier scores of the methods do not say much as They lay close to each other. In this case, Combining two similarity methods does not lead to a better result as well.

These results are not expected, considering the Resnik and Lin semantic similarity methods are specific for within an ontology and the MCS similarity method can be used for all graphs in general. This could mean multiple things. For example, the graph structure could contain more information about the abnormalities than the semantic content. This is one of the areas, where future research could expand upon. The prediction of other symptoms and abnormalities can be researched in the future as well. It would be interesting to know how this classifier would handle the prediction of other symptoms. The usage of various similarity methods could also be extended.

Three methods have been tested in this research, but there are far more methods available, especially in the graph similarity domain. Additional research could also give more insight into the difference of importance between graph structure and semantic information.

5.1 Limitations

The research uses a small data set, which contains only 553 patients. Furthermore, the dataset only contains a few select syndromes, so it does not represent the average patient. This means that it is not possible to generalise these results at this moment, until more data are collected. Something to also take into consideration is the fact that in this research, the simulation of the 'new' patients only removes the nodes and descendants from heart and kidney abnormality. What happens if there is a patient with only a few symptoms? This is a question, which could be answered in future research. For now, this classifier can be used to determine the order in which patients will undergo imaging in the hospital. It can also be used to make a well thought-out consideration in developing countries, where imaging machines are sparse and these choices should be made wisely. Moreover, it could be used in a crisis situation, when there is a need of imaging.

5.2 Conclusion

In this study, it was researched which graph or semantic similarity method works best when predicting whether patients have heart or kidney abnormalities based on their phenotype. The research shows that the maximum common sub-graph similarity method yields the best result when predicting this. In both the heart abnormality and kidney abnormality cases, the classifier that uses MCS leads to the highest AUC score. The MCS method is also able to achieve the best balance between FPR and TPR of the three methods. The classifier can be fine-tuned to the liking of the clinician to achieve this best balance or a TPR of 1. The classifier performs quite well and could already be used for some tasks, which do not require a TPR of 1, such as the determination of the order in which patients undergo imaging. In future research, the usage of a larger data set could improve the generalizability of the study.

Chapter 6

Data Availability

The code used in this research is available at: https://github.com/MisjaZippelius/BachelorThesis
The used dataset is however not publicly available, because the data might be traceable and would violate the General Data Protection Regulation restrictions.

Bibliography

[1] Godwin Inalegwu Ogbole, Adekunle Olakunle Adeyomoye, Augustina Badu-Peprah, Yaw Mensah, and Donald Amasike Nzeh. Survey of magnetic resonance imaging availability in west africa. *Pan African Medical Journal*, 30(1), 2018.

[2] Marianna Milano. Gene prioritization tools. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 907–914. Academic Press, Oxford, 2019.

[3] Lada A. Adamic. *World Wide Web, Graph Structure*, pages 10058–10072. Springer New York, New York, NY, 2009.

[4] Yves Robert. *Task Graph Scheduling*, pages 2013–2025. Springer US, Boston, MA, 2011.

[5] David F. Nettleton. Data mining of social networks represented as graphs. *Computer Science Review*, 7:1–34, 2013.

[6] Wei-Nchih Lee, Nigam Shah, Karanjot Sundlass, and Mark Musen. Comparison of ontology-based semantic-similarity measures. In *AMIA annual symposium proceedings*, volume 2008, page 384. American Medical Informatics Association, 2008.

[7] Elena Rojano, José Córdoba-Caballero, Fernando M Jabato, Diana Gallego, Mercedes Serrano, Belén Pérez, Álvaro Parés-Aguilar, James R Perkins, Juan AG Ranea, and Pedro Seoane-Zonjic. Evaluating, filtering and clustering genetic disease cohorts based on human phenotype ontology data with cohort analyzer. *Journal of personalized medicine*, 11(8):730, 2021.

[8] Mohammed Diykh, Yan Li, and Peng Wen. Eeg sleep stages classification based on time domain features and structural graph similarity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(11):1159–1168, 2016.

[9] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554, 2019.

[10] S Vijayarani, S Dhayanand, and M Phil. Kidney disease prediction using svm and ann algorithms. *International Journal of Computing and Business Research (IJCBR)*, 6(2):1–12, 2015.

[11] What is a graph (data structure)? `https://www.educative.io/edpresso/what-is-a-graph-data-structure`. Accessed: 2021-12-31.

[12] Li Zeng, Yan Jiang, Weixin Lu, and Lei Zou. Deep analysis on subgraph isomorphism, 2021.

[13] Graph theory - isomorphism. `https://www.tutorialspoint.com/graph_theory/graph_theory_isomorphism.htm`. Accessed: 2022-01-33.

[14] Laura Zager. *Graph similarity and matching.* PhD thesis, Massachusetts Institute of Technology, 2005.

[15] Ruud Welling. A performance analysis on maximal common subgraph algorithms. In *15th Twente Student Conference on IT, Enschede, The Netherlands. University of Twente*, volume 14, 2011.

[16] What is semantic similarity. `https://www.igi-global.com/dictionary/semantic-similarity/26364`. Accessed: 2021-12-31.

[17] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130, 1999.

[18] Dekang Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304, 1998.

[19] Kevin Arvai, Carlos Borroto, Vladimir Gainullin, and Kyle Retterer. Phenopy, 2019.

[20] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[21] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, et al. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1):D1207–D1217, 2021.

[22] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[23] Kaspar Rufibach. Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–939, 2010.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.