Similarity Measure Development for Case-Based Reasoning–A Data-Driven Approach

Deepika Verma[1](✉), Kerstin Bach[1], and Paul Jarle Mork[2]

[1] Department of Computer Science,
Norwegian University of Science and Technology, Trondheim, Norway
deepika.verma@ntnu.no
[2] Department of Public Health and Nursing,
Norwegian University of Science and Technology, Trondheim, Norway
http://www.idi.ntnu.no, http://www.ntnu.no/ism

Abstract. In this paper, we demonstrate a data-driven methodology for modelling the local similarity measures of various attributes in a dataset. We analyse the spread in the numerical attributes and estimate their distribution using polynomial function to showcase an approach for deriving strong initial value ranges of numerical attributes and use a non-overlapping distribution for categorical attributes such that the entire similarity range [0,1] is utilized. We use an open source dataset for demonstrating modelling and development of the similarity measures and will present a case-based reasoning (CBR) system that can be used to search for the most relevant similar cases.

Keywords: Case-based reasoning · Local similarity modelling · Knowledge modelling

1 Introduction

CBR has gained popularity in the recent years due to its novel approach to abstract and transfer domain-specific expert knowledge into a user-friendly tool which offers appropriate reasoning for solutions to problems ranging from simple daily life tasks to complex tasks which otherwise necessitate expert guidance.

Modelling the local similarities of attributes while preparing a CBR model can be a challenging task for small and simple, and large and complex data sets alike. In this paper, we direct our attention towards the knowledge engineering process of creating a CBR model and present a data-driven approach for modelling local similarity measures using the openly available User Knowledge Modelling dataset[1] in the myCBR workbench [2,6]. The main contribution of this paper is a methodology for modelling the local similarity measures using a data-driven approach. We will showcase how the knowledge stored in a data set can be leveraged to define strong initial value ranges for both numerical and categorical attributes and therewith moderate and stratify the knowledge modelling process.

[1] https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling.

K. Bach and M. Ruocco (Eds.): NAIS 2019, CCIS 1056, pp. 143–148, 2019.
https://doi.org/10.1007/978-3-030-35664-4_14

The remainder of this paper is organised into sections as follows: in Sect. 2, we discuss related work about the use of data-driven similarity measure development and its application in CBR, followed by Sect. 3 wherein we present our similarity modelling approach. Finally, Sect. 4 concludes the work presented in this paper.

2 Related Work

Similar to the preference-based similarity measure development framework presented by authors in [1,4], we are presenting a framework for modelling local similarity measures based on the data set available. Therewith we can tailor each similarity measure to the application domain. Using a data-driven approach for automatic similarity learning and feature weighting has been presented by Gabel and Godehardt [3] where they trained a neural network to induce local and global similarity measures [5]. While we are not automatically assigning the similarity measures, we use the existing cases to derive them.

3 Data-Driven Knowledge Modelling

In this section, we explain how we implement a CBR system that can be applied to find the most similar and relevant cases. We use the local-global-principle [5] for tailoring the similarity measure for each attribute and thereby build a knowledge model. Once the local similarity measures are defined, we continue to use weighted sum for defining the global similarity.

Some of the most common challenges for utilizing any dataset for developing a CBR system are the identification of suitable dataset context for the problem at hand, definition of initial similarity measures, representation of cases and determination of valuable cases for populating the case base. In this section, we first describe how we populate the case base and generate cases in the developed case representation. Then we present our method for utilizing a given dataset to model the local similarity measures for both numerical as well as categorical attributes.

3.1 Case Generation

Developing a case representation is the first step of the CBR system development. Depending on the domain and the available data this can be a challenging process on its own. For presenting our data-driven modelling technique, we use the User Knowledge Modelling dataset, which comprises of six attributes, five numerical and one categorical. The description of all the attributes is presented in Table 1.

The categorical attribute *USN* has four permitted values: *Very Low, Low, Middle, High*. Table 2 shows the data statistics of the numerical attributes in the dataset.

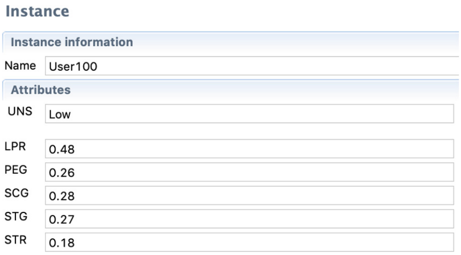
The case base is then populated by loading the dataset into the previously defined case representation in the myCBR workbench. A single case in myCBR is represented as shown in Fig. 1, where *User* is the name of the concept which comprises of six attributes present in the original dataset.

Table 1. Description of attributes in User Knowledge Modelling dataset

Attribute	Description
STG	The degree of study time for goal object materials
SCG	The degree of repetition number of user for goal object materials
STR	The degree of study time of user for related objects with goal object
LPR	The exam performance of user for related objects with goal object
PEG	The exam performance of user for goal objects
UNS	The knowledge level of user

Table 2. Data set statistics

	STG	SCG	STR	LPR	PEG
count	403	403	403	403	403
mean	0.3531	0.3559	0.4576	0.4313	0.4563
min	0	0	0	0	0
max	0.99	0.90	0.95	0.99	0.99



Instance

Instance information

Name: User100

Attributes

UNS: Low

LPR: 0.48

PEG: 0.26

SCG: 0.28

STG: 0.27

STR: 0.18

Fig. 1. Case representation in myCBR

3.2 Data-Driven Similarity Measures Development

The local-global-principle requires both the local similarity measure on the attribute level and the global one on the conceptual to be defined.

Researchers in CBR domain face the challenge of balancing the input from the domain experts and the available data while modelling the local similarity measures for different attributes in myCBR. Having a criteria which can lead the knowledge modelling process is helpful for both parties. We therefore suggest to make use of the existing data in this process. While setting upper and lower limits for numerical attributes is straight-forward, assigning the similarity behaviour is not. Consecutively, we assume that local similarity measures for continuous numerical attributes are polynomial distance functions (due to their flexibility and better converging ability) and the question is how steep of a similarity decline should be chosen. Therefore, we focus on the polynomial function of the similarity measure for numerical attributes and our goal is to determine their degree. We use box plots for visualizing the distributions and variations in the data set and map this into modelling local similarity measures.

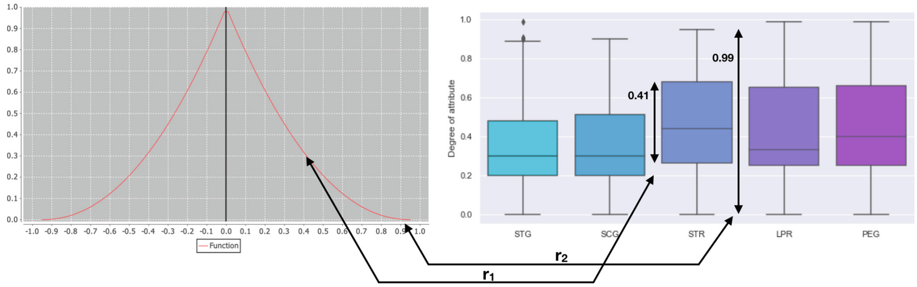


Fig. 2. Example for data-driven local similarity modelling: on the left there is a screen shot of a polynomial similarity function for a value range between 0 and 1. With the arrows we depict how the box-plot for attribute STR relates to the decrease in similarity at a certain distance.

Figure 2 shows an example of a local similarity measure for a numerical attribute. From there we look into the Q_1 and Q_3, which indicate the majority spread of the attributes in the data set. In line with [1,7], we decided to take these values as reference points for determining the decrease in similarity.

Hence, creating a box-plot of the data set will allow modelling each attribute since we only take the Inter Quartile Range (IQR) and the range (min to max) into account:

$$\begin{aligned} r_1 &= IQR \\ r_2 &= range \end{aligned} \tag{1}$$

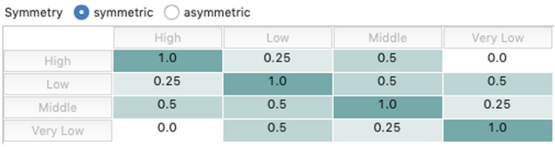
It represents the difference between upper (Q_3) and lower (Q_1) quartiles in the box-plot, that is $IQR = Q_3 - Q_1$.

We assume that all similarity functions are polynomial and adjust the polynomial degree of the similarity function such that

$$\begin{aligned} y(r_1) &\approx 0.30 \\ y(r_2) &\approx 0 \end{aligned} \tag{2}$$

We can observe in Fig. 2 how the similarity function varies with respect to the attribute value after applying the methodology in Eqs. 1 and 2. The bigger the polynomial degree, the steeper the similarity function and more precise the attribute values in retrieved cases. The decline in the similarity function is steeper in the beginning until at r_1 it reaches close to $y(r_1)$ and then decreases gradually until at r_2 it is approximately close to $y(r_2)$. This way, the similarity function covers the entire attribute range as well as the similarity measure range $[0, 1]$. We use this as the initial definition of similarity measures.

While the local similarity measures for numerical attributes can be derived using their data distributions, assigning the similarity behaviour for categorical attributes can be challenging as it depends on whether or not there is a pre-existing relationship between the categorical values. In our dataset, the categorical attribute *UNS* has four permitted values which have an implicit relationship amongst each other. The local similarity measure for such an attribute can be modelled such that the relationship amongst the values is preserved while achieving the desired variation in the similarity measure in the range [0,1], as shown in Fig. 3. In case of no relationship amongst the values, the similarity of one value to every different value can be set to zero.



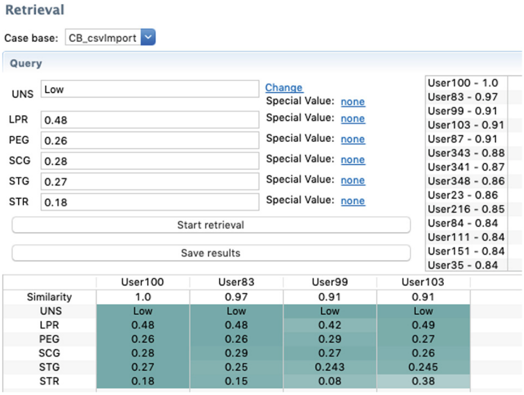
Symmetry ● symmetric ○ asymmetric

	High	Low	Middle	Very Low
High	1.0	0.25	0.5	0.0
Low	0.25	1.0	0.5	0.5
Middle	0.5	0.5	1.0	0.25
Very Low	0.0	0.5	0.25	1.0

Fig. 3. Similarity measure modelling for non-overlapping categorical attribute

3.3 Retrieving Similar Cases

Once the casebase and similarity measures are in place, the model can be used to find similar cases. Figure 4 shows the result of one such query retrieval in myCBR. The retrieved cases are sorted by similarity value in descending order, that is, most similar case are displayed at the top while least similar are at the bottom. On the lower part of the figure, the four most similar *Users* are shown in a detailed view. The tool marks closer matches darker.



Case base: CB_csvImport

UNS	Low	Special Value: none
LPR	0.48	Special Value: none
PEG	0.26	Special Value: none
SCG	0.28	Special Value: none
STG	0.27	Special Value: none
STR	0.18	Special Value: none

User100 - 1.0
User83 - 0.97
User99 - 0.91
User103 - 0.91
User87 - 0.91
User343 - 0.88
User341 - 0.87
User348 - 0.86
User23 - 0.86
User216 - 0.85
User84 - 0.84
User111 - 0.84
User151 - 0.84
User35 - 0.84

	User100	User83	User99	User103
Similarity	1.0	0.97	0.91	0.91
UNS	Low	Low	Low	Low
LPR	0.48	0.48	0.42	0.49
PEG	0.26	0.26	0.29	0.27
SCG	0.28	0.29	0.27	0.26
STG	0.27	0.25	0.243	0.245
STR	0.18	0.15	0.08	0.38

Fig. 4. A query and its retrieval result in the myCBR workbench

4 Discussion and Conclusion

In this paper, we have presented an approach to model the local similarity measures of a given dataset in myCBR in a data-driven manner. Our approach can be applied on any dataset to model the similarity measures. A more detailed evaluation of our approach can be found in [7] where we statistically evaluated its effectiveness using a public health domain dataset and showed that the CBR model created using our approach outperforms the k-NN regressor model in finding the most similar cases. The approach presented in this work can significantly reduce the efforts required to create new CBR models using different data sets from scratch. Therefore, it is safe to conclude that the approach works well on the used dataset and may also be applicable to other domains.

References

1. Abdel-Aziz, A., Strickert, M., Hüllermeier, E.: Learning solution similarity in preference-based CBR. In: Lamontagne, L., Plaza, E. (eds.) ICCBR 2014. LNCS (LNAI), vol. 8765, pp. 17–31. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11209-1_3
2. Bach, K., Althoff, K.-D.: Developing case-based reasoning applications using myCBR 3. In: Agudo, B.D., Watson, I. (eds.) ICCBR 2012. LNCS (LNAI), vol. 7466, pp. 17–31. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32986-9_4
3. Gabel, T., Godehardt, E.: Top-down induction of similarity measures using similarity clouds. In: Hüllermeier, E., Minor, M. (eds.) ICCBR 2015. LNCS (LNAI), vol. 9343, pp. 149–164. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24586-7_11
4. Hüllermeier, E., Schlegel, P.: Preference-based CBR: first steps toward a methodological framework. In: Ram, A., Wiratunga, N. (eds.) ICCBR 2011. LNCS (LNAI), vol. 6880, pp. 77–91. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23291-6_8
5. Richter, M.M.: The knowledge contained in similarity measures. In: Veloso, M.M., Aamodt, A. (eds.) ICCBR-1995. LNCS, vol. 1010. Springer, Heidelberg (1995)
6. Stahl, A., Roth-Berghofer, T.R.: Rapid prototyping of CBR applications with the open source tool myCBR. In: Althoff, K.-D., Bergmann, R., Minor, M., Hanft, A. (eds.) ECCBR 2008. LNCS (LNAI), vol. 5239, pp. 615–629. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85502-6_42
7. Verma, D., Bach, K., Mork, P.J.: Modelling similarity for comparing physical activity profiles - a data-driven approach. In: Cox, M.T., Funk, P., Begum, S. (eds.) ICCBR 2018. LNCS (LNAI), vol. 11156, pp. 415–430. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01081-2_28