L. Spitz*, S. C. Umeh, D. Behme, B. Neyazi, I. E. Sandalcioglu, B. Preim, and S. Saalfeld

Subset selection for intracranial aneurysms for training simulations

https://doi.org/10.1515/cdbme-2024-1019

Abstract: We present a framework with two options for selecting a subgroup of training cases for intracranial aneurysm (IA) treatment. Option one, general training, describes training cases that represent a variety of different IAs that represent the diversity of real world cases. This can be achieved via instance selection (IS), which reduces size of a dataset by eliminating redundancies via outlier removal and clustering. Option two, a specific training scenario, describes training that is specialized based on IA features of a specific case, for which we present the novel reverse instance selection (RIS), which introduces similarity to the specific case to the IS methodology.

We evaluated our IS and RIS by comparing them to subsets selected based on similarity (SIM) and random sampling (RndS). RIS outperformed SIM and RndS in 79% of experiments. IS outperformed RndS in 33% of experiments. In both scenarios, we observed that our approaches, which balance the weaknesses of SIM and RndS, perform best for small subset sizes close to the database cluster size.

Our IS and RIS are flexible in regards to the underlying machine learning and weighting of metrics for evaluation, thus providing a way to select a representative and diverse subset not just for IAs, but also for different kinds of data.

Keywords: intracranial aneurysms, instance selection, subgroup selection

1 Introduction

Intracranial aneurysms (IAs) are pathologic dilatations on neurovascular vessel walls that bear the risk of rupture. Various features, including morphologic parameters describing their shapes and hemodynamic parameters describing the blood flow within them, as well as metadata referring to the patient-related characteristics, are relevant in IA research, rupture assessment, and determining treatment options [1, 2]. Treatments include endovascular interventions like coiling or stenting, and open surgery, where the skull is opened and a clip is inserted to stall blood flow into the IA sack. Due to lower complications, faster recovery times and adequate occlusion rates, endovascular treatment is becoming more and more common, but complex cases still need to be surgically clipped [3].

Training simulations enable trainees to gather experience, gain spatial skills, and learn the procedure of an intervention in a safe environment. Especially for surgical treatment, where there are less cases to train on, training simulations can be crucial [4]. As more and more training simulations are developed, the question of training case selection arises. The IAs a student trains on should be realistic and representative of the scenarios they may encounter in real patients.

We present a framework for the selection of a subgroup of training cases for IA treatment. In this framework there are two options for training case selection: general training, and training for a specific scenario. General training describes training cases that represent a variety of different IAs that a physician may need to treat. A specific training scenario describes training that is specialized based on IA features of a specific case.

For general training, instance selection (IS) can be utilized. IS is the process of reducing the size of a dataset while preserving quality and variance. Thus it can be used to select a subset of representative and diverse IAs from a large database of IAs. IS can be done in a variety of ways, popular models including prototype selection, training set selection, and sampling, clustering and prototyping [5, 6]. In the medical context IS has also been used, for example via clustering approaches for medical relations [7], or in a divide-and-conquer based strategy to tackle the specific issues of medical datasets [8].

For training for specific cases, we introduce reverse instance selection (RIS), a novel approach where a representative and diverse subset is chosen with respect to one specified IA. Here, the subset shall be similar to the specified IA, without loosing the variance of the entire database.

2 Materials & Methods

The framework (overview in Fig. 1) was implemented in Python 3.10.12. We had access to two databases in the form

***Corresponding author: L. Spitz, S. C. Umeh, B. Preim,** Department of Simulation and Graphics, Otto-von-Guericke University, Magdeburg, Germany, e-mail: spitz@isg.cs.ovgu.de
***Corresponding author: L. Spitz, D. Behme, B. Preim, S. Saalfeld,** Forschungscampus STIMULATE, Magdeburg, Germany
D. Behme, Department of Neuroradiology, University Hospital Magdeburg, Germany
B. Neyazi, I. E. Sandalcioglu, Department of Neurosurgery, University Hospital Magdeburg, Germany
S. Saalfeld, Department of Medical Informatics, University Hospital Schleswig-Holstein, Campus Kiel, Germany

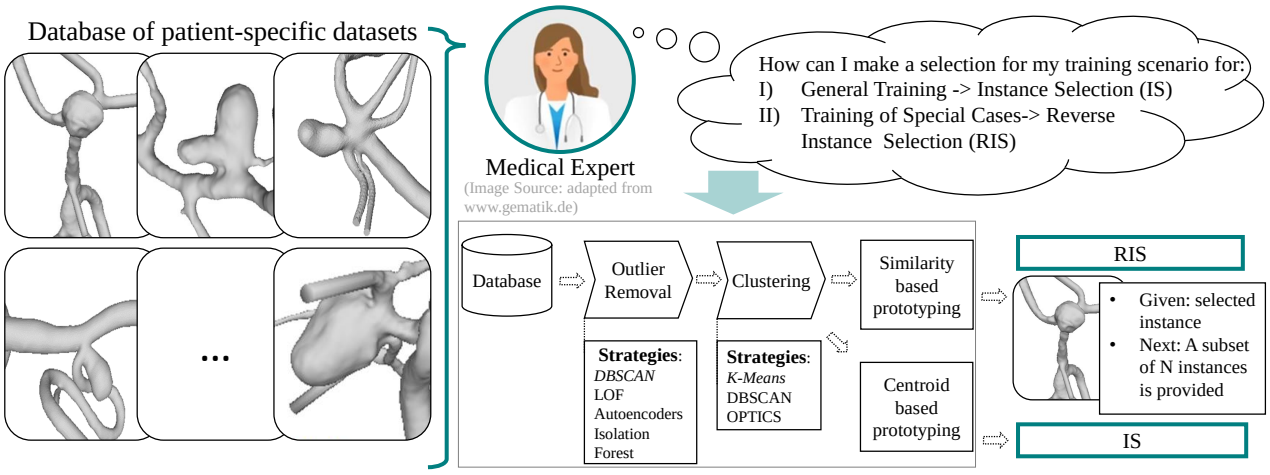

Fig. 1: Overview over subset selection for training, including out framework for RIS and IS.

of tabular data. The first database, for which we also extracted 3D models, had 76 IAs from 54 patients collected from three clinics, with a total of 37 parameters (7 metadata, 23 morphological [9], 7 hemodynamic [10]). After preprocessing, 70 IAs with 29 features remained. The second database had 406 IAs with 106 parameters, though after preprocessing, 351 IAs and 12 parameters (metadata and morphological) remained. During preprocessing, duplicate records or features with very high correlation were removed, as well as entries with missing data. Categorical and boolean features were encoded numerically, and all features were scaled to have a mean of 0 and a standard deviation of 1. Principle component analysis (PCA) served as dimension reduction, and the first three principal components were used to continue calculations, as they covered a sufficient explained variance of at least 88%.

We determined three criteria necessary for subgroup selection: inlierness, diversity, and, for RIS, similarity.

The inlierness criterion ensures that there are no outliers in the selected subset. It is fulfilled via outlier removal. We tested density-based spatial clustering of applications with noise (DBSCAN) [11], local outlier factor (LOF) [12], autoencoders [13] and isolation forest [14]. DBSCAN ($minPts = 4, eps = 1.6$, determined empirically) was the best candidate as it requires few hyperparameters, no input number of outliers like autoencoders, and no clearly defined clusters like LOF.

The diversity criterion ensures variance and low redundancy in the selected subset. It is fulfilled via clustering. We tested DBSCAN, ordering point to identify clustering structure (OPTICS) [15], and k-means [16]. The latter was selected for our application as it showed similar results for our database with and without PCA.

Prototyping selects the candidates in the subgroup, and in case of RIS meets the similarity criterion for the specified IA. It describes an iterative process of choosing candidates from the clusters. For RIS, the candidates are chosen based on how similar they are to the specified case. For IS, the candidates closest to the centroids of the clusters are chosen.

As we present RIS as a new concept, we also introduce its evaluation score which is made up of three weighted metrics for a selected subset: similarity, redundancy, and anomaly. Similarity ensures representation of the specified IA case in the selected subset . One part of the similarity metric thus calculates how similar two aneurysms are based on their parameters or PCs (denoted as vectors p, q), utilizing Euclidean distance $dist(p, q)$ and some modifications:

$$sim(p, q) = \frac{1}{1 + dist(p, q)} \tag{1}$$

Similarity for an entire subset thus is the mean similarity of all IAs in the subset S to the specified case A:

$$Sim(A, S) = \frac{\Sigma_{s_i \epsilon S} sim(A, s_i)}{|S|} \tag{2}$$

Redundancy assures diversity by punishing similarity within the selected subset. It can be calculated using Eq. 1, and thus the redundancy of one IA s_i to an extracted subset S is:

$$Red(s_i, S) = 1 - \frac{\Sigma_{s_j \epsilon S} sim(s_i, s_j)}{|S| - 1} \tag{3}$$

The overall redundancy of one subset is therefore:

$$Red(S) = \frac{\Sigma_{s_i \epsilon S} Red(s_i, S)}{|S|} \tag{4}$$

Anomaly, or outlierness, of an IA punishes inclusion of outliers in a subset. It can be calculated using local reachability distance (LRD) which is local density based on neighbouring instances within the database. LRD is defined by:

$$LRD_k(s_i) = \frac{1}{\frac{1}{k}\Sigma_{q \epsilon N_k} reachDist_k(s_i, q)} \tag{5}$$

where $N_k(s_i)$ are the k nearest neighbours to IA s_i, and $reachDist_k(p,q)$ is the reachability distance, which is the maximum of the distance between two points and the distance to the point's kth neighbour. For overall anomaly of a subset, the LDR of each point is normalized and the average is taken:

$$Anom_k(S) = \frac{\Sigma_{s_i \epsilon S} norm(LRD_k(s_i))}{|S|} \quad (6)$$

The value of k can be adjusted based on individual use case, but for our case we empirically determined and set k to four.

Similarity, redundancy and anomaly metrics can then be combined into an overall score, where each metric can additionally be weighted based on domain knowledge or use case.

$$Score(S) = w_1 Sim(A,S) + w_2 Red(S) + w_3 Anom_k(S) \quad (7)$$

As we are working on a sparse dataset with a focus on representation and diversity, we set the weights to favour similarity and redundancy over anomaly: $w_1 = 0.4, w_2 = 0.4, w_3 = 0.2$.

3 Results

We compared our RIS approach to subsets consisting of the most similar IAs to the specified IA (SIM), and to randomly sampled (RndS) subsets. We tested with six randomly selected specified IAs, and with two subset sizes, one equalling the number of clusters in the dataset ($N = 3$) and a bigger one ($N = 5$) (Tab. 1).

Tab. 1: Results of comparing RIS with SIM and RndS models. Metrics are averages over six experiments with randomized specified IAs. Best performing models and metrics are highlighted in bold.

Database	Model	Metrics			
		Sim	Red	$Anom$	$Score$
1	**RIS**	0.399	0.744	0.252	**0.508**
($N = 3$)	Similarity	**0.568**	0.446	**0.449**	0.506
	Random	0.225	**0.809**	0.234	0.455
1	RIS	0.392	0.715	0.239	0.495
($N = 5$)	**Similarity**	**0.529**	0.500	**0.432**	**0.505**
	Random	0.245	**0.777**	0.276	0.463
2	**RIS**	0.397	0.701	**0.638**	**0.567**
($N = 3$)	Similarity	**0.528**	0.485	0.595	0.524
	Random	0.201	**0.832**	0.632	0.540
2	**RIS**	0.385	0.694	0.622	**0.556**
($N = 5$)	Similarity	**0.486**	0.552	0.610	0.537
	Random	0.138	**0.814**	**0.630**	0.531

RIS outperformed SIM and RndS subsets in 79% of experiments, though SIM also performed well, as visible in the mean results of Tab. 1. In the evaluation metrics, SIM performed best for the Sim metric in all experiments, and RndS best for Red metric in all but one experiments. Results for $Anom$ metric were more scattered, SIM performed best in 50%, RndS best in 29% and RIS best in 21% of experiments. While RIS did not outperform in any one metric, it aims to strike a balance between the weaknesses of SIM and RndS. However, for both datasets, we observed that when subset size exceeded cluster size within the database, RIS performance grew worse, and SIM subset selection performed better.

To evaluate our IS approach, we used the common IS evaluation method of comparing classification accuracy of a model trained on the complete dataset and on the extracted subset [8]. Using XGBoost [17] and a 80/20 training/testing split, we trained several models with different subset sizes (based on % of training data) and compared IS subsets results to those of RndS subsets. For each subset size, five experiments were done. We utilized database 2 and the Iris dataset [18, 19]. Classification accuracy of the model trained on the complete dataset was 100%.

Tab. 2: Averages of results of comparing accuracy of a classification model trained on subsets extracted with IS and random sampling, tested on two databases. Subset sizes are % of the complete dataset.

Database	Model	Accuracy per subset size (%)		
		12.5%	30%	50%
Iris	RIS	**83.3**	**96.7**	96.7
	Random	77.3	94.7	**99.3**
2	RIS	45.1	53.5	54.9
	Random	**49.9**	**56.3**	**56.6**

For subset sizes 12.5% and 30% of the IRIS dataset, RIS outperformed the RndS selection. For subset size of 50% and all sizes of database 2, RndS achieved a higher accuracy than RIS. However, for database 2, RndS accuracy did not get higher than 57%, and differences to RIS were always smaller than 5%.

4 Discussion & Conclusion

The results in Tables 1 and 2 show that our presented approach for RIS and IS performs best for small subset sizes close or equal to the number of ideal clusters in the database. In case of IS, where RndS achieved better results, database 2, even when sampled at 12.5%, may simply be too big to emphasize the strengths of the presented IS approach. Even so, RndS only performed 5% better at most.

Our IS approach and our novel method of RIS can easily be adapted in parameters and are flexible in which machine learning techniques it utilizes in the anomaly detection, clustering and prototyping steps. They can thus be adapted to many different datasets if domain knowledge is available. The same flexibility is applicable to our presented evaluation method for RIS, where the weights of each metric can be adjusted based on which metric is more favoured.

The currently chosen implementation of the approach has limitations in that is is based on distance metrics and needs a minimum subset size equal to the numbers of clusters. The latter may be problematic for databases with a lot of clusters, but can be solved by using another clustering method, or by not utilizing all clusters through weighting or rating clusters before prototyping. Distance metrics could also be replaced by other metrics.

The adjustment of weighting the metrics in our evaluation is an advantage due to adaptability, but this also meant no ideal weights were determined. A dimension reduction method that can handle categorical features without numerical encoding like we did for PCA would be desirable in the future.

We discussed the IS and RIS approach with clinical experts, which highlighted the importance of adequate IS and which were interested in using the tool to quickly and automatically identify initial subsets. However, we could not conduct a qualitative user study due to a missing ground truth which is difficult to determine and additionally highly dependent on each medical application scenario.

In this work, we present an approach for IS, and the novel concept and approach of RIS and a metric score for its evaluation, for training simulations of IA treatment. Both IS and RIS are useful for selecting a subset of training cases for IA treatment. RIS can select a subset of IAs similar to one specified case that shall be trained for in particular, whereas IS selects a diverse and representative subset of the entire database, so trainees can gain proficiency in a variety of IAs.

Author Statement

This work is partially funded by the Federal Ministry of Education and Research within the Forschungscampus STIMULATE (grant no. 13GW0473A) and the German Research Foundation (SA 3461/3-1, SPP2311, Nr. 465189657). The authors state no conflict of interest. Informed consent was obtained from all individuals included in this study. The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Dhar S, Tremmel M, Mocco J, Kim M, Yamamoto, J, Siddiqui AH, et al. Morphology parameters for intracranial aneurysm rupture risk assessment. Neurosurgery 2008, 63(2):185–197.

[2] Han P, Jin D, Wei W, Song C, Leng X, Liu L, et al. The prognostic effects of hemodynamic parameters on rupture of intracranial aneurysm: A systematic review and meta-analysis. Int J Surg 2021, 86:15-23.

[3] Lin N, Cahill K, Frerichs K, Friedlander R, Claus E. Treatment of ruptured and unruptured cerebral aneurysms in the USA: a paradigm shift. J Neuro Interv Surg 2018, 10:i69–i76.

[4] Joseph FJ, Vanluchene HER, Bervini D. Simulation training approaches in intracranial aneurysm surgery—a systematic review. Neurosurg Rev 2023, 46(101).

[5] Reinartz TA. Unifying View on Instance Selection. Data Min Knowl Discov 2002, 6:191–210.

[6] García S, Luengo J, Herrera F. Data preprocessing in data mining. Springer, 2015.

[7] Kim Y, Riloff E, Meystre SM. Exploiting Unlabeled Texts with Clustering-based Instance Selection for Medical Relation Classification. AMIA Annu Symp Proc 2018, 2017:1060-1069.

[8] Huang M-W, Chih-Fong T, Wei-Chao L. Instance selection in medical datasets: A divide-and-conquer framework. Comput Electr Eng 2021, 90.

[9] Saalfeld S, Berg P, Niemann A, Luz M, Preim B, Beuing O. Semiautomatic neck curve reconstruction for intracranial aneurysm rupture risk assessment based on morphological parameters. Int J Comput Assist Radiol Surg 2018, 13(11):1781–93.

[10] Cebral JR, Mut F, Weir J, Putman CM. Association of hemodynamic characteristics and cerebral aneurysm rupture. AJNR Am J Neuroradiol 2011, 32:264–70.

[11] Ester M, Kriegel H-P, Sander J, Xu X: A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. of 2nd KDD 1996, p.226–231.

[12] Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: Identifying Density-based Local Outliers. ACM SIGMOD Record 2000, p. 29(2):93–104.

[13] Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal 1991, 37(2):233–243.

[14] Liu FT, Ting KM, Zhou Z-H. Isolation Forest. Proc. of 8th ICDM, 2008, p.413-422.

[15] Ankerst M, Breunig MM, Kriegel H-P, und Sander J. OPTICS: Ordering points to identify the clustering structure. ACM SIGMOD Record 1999, 28(2):49–60.

[16] MacQueen J. Some methods for classification and analysis of multivariate observations. Proc. of 5th Berkeley Symp Math Statist Probability 1967, p.281–297.

[17] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc. of 22nd ACM SIGKDD 2016, p.785–94.

[18] Anderson E. The species problem in iris. Ann Mo Bot Gard 1936, 23(3):457-509.

[19] Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen 1936. 7(2):179–188.