EuroVis Workshop on Visual Analytics (2019)
C. Turkay and T. von Landesberger (Editors)

Visual Analysis of Degree-of-Interest Functions to Support Selection Strategies for Instance Labeling – Supplemental Materials

Jürgen Bernard [1], Marco Hutter[1], Christian Ritter [1], Markus Lehmann[1], Michael Sedlmair[2], and Matthias Zeppelzauer[3]

[1]TU Darmstadt, Darmstadt, Germany
[2] University of Stuttgart, Stuttgart, Germany
[3]St. Pölten University of Applied Sciences, Austria

Abstract

In addition to the manuscript, the supplemental materials document contains to two tables with details about our taxonomy of DOI (degree-of-interest) functions. The overal taxonomy is split into two parts by the primary distinction criterion, i.e., data-based and model-based DOIs. Both tables in this document (for data-based an model-based DOIs) contain more details about sub-categories of the taxonomy and references to techniques and implementations. Along these lines, a third level of depth is introduced reflecting important leaves of the hierarchy, i.e., concrete DOIs. This hierarchy level is encoded with standard font, whereas the inner branches of the taxonomy are dyed bold.

References

[BKNS00] BREUNIG M. M., KRIEGEL H.-P., NG R. T., SANDER J.: Lof: Identifying density-based local outliers. In *Int. Conf. On Management of Data (SIGMOD)* (2000), ACM. 2

[BSB*15] BERNARD J., SESSLER D., BANNACH A., MAY T., KOHLHAMMER J.: A visual active learning system for the assessment of patient well-being in prostate cancer research. In *IEEE VIS WS on Visual Analytics in Healthcare (VAHC)* (2015), ACM, pp. 1–8. doi:10.1145/2836034.2836035. 2

[BZL*18] BERNARD J., ZEPPELZAUER M., LEHMANN M., MÜLLER M., SEDLMAIR M.: Towards User-Centered Active Learning Algorithms. *Computer Graphics Forum (CGF)* (2018). doi:10.1111/cgf.13406. 2, 3

[CBK09] CHANDOLA V., BANERJEE A., KUMAR V.: Anomaly detection: A survey. *ACM Comput. Surv. 41*, 3 (2009), 15:1–15:58. doi:10.1145/1541880.1541882. 2

[Dun74] DUNN J. C.: Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems 4*, 1 (1974), 95–104. 2, 3

[FT04] FUGLEDE B., TOPSOE F.: Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Thsome refeory, 2004. ISIT 2004. Proceedings.* (2004). doi:10.1109/ISIT.2004.1365067. 3

[HBV02] HALKIDI M., BATISTAKIS Y., VAZIRGIANNIS M.: Clustering validity checking methods: Part ii. *SIGMOD Rec. 31*, 3 (2002), 19–27. doi:10.1145/601858.601862. 2

[Jai10] JAIN A. K.: Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31*, 8 (2010), 651–666. 2

[KL*51] KULLBACK S., LEIBLER R., ET AL.: On information and sufficiency. *The Annals of Mathematical Statistics 22*, 1 (1951), 79–86. 3

[KN98] KNORR E. M., NG R. T.: Algorithms for mining distance-based outliers in large datasets. In *Conference on Very Large Data Bases (VLDB)* (1998), Morgan Kaufmann Publishers Inc., pp. 392–403. URL: http://dl.acm.org/citation.cfm?id=645924.671334. 2

[Kol33] KOLMOGOROV A.: Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari 4* (1933), 83–91. 3

[Mam98] MAMITSUKA N. A. H.: Query learning strategies using boosting and bagging. In *International Conference on Machine Learning (ICML)* (1998), vol. 1, Morgan Kaufmann Pub. 3

[Rou87] ROUSSEEUW P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20* (1987), 53 – 65. doi:https://doi.org/10.1016/0377-0427(87)90125-7. 2, 3

[RRS00] RAMASWAMY S., RASTOGI R., SHIM K.: Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec. 29*, 2 (2000), 427–438. doi:10.1145/335191.335437. 2

[SC08] SETTLES B., CRAVEN M.: An analysis of active learning strategies for sequence labeling tasks. In *Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA, USA, 2008), Computational Linguistics, pp. 1070–1079. 3

[Set12] SETTLES B.: Active learning. *Synthesis Lectures on Artif. Intell. and Machine Learning 6*, 1 (2012), 1–114. 3

[Sha48] SHANNON C. E.: A mathematical theory of communication. *The Bell System Technical Journal 27*, 3 (1948), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x. 3

[Sim49] SIMPSON E. H.: Measurement of diversity. *Nature 163*, 4148 (1949), 688. 3

[Smi48] SMIRNOV N.: Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics 19*, 2 (1948), 279–281. 3

[SOS92] SEUNG H. S., OPPER M., SOMPOLINSKY H.: Query by committee. In *Worksh. on Comput. Learning Theory (COLT)* (New York, NY, USA, 1992), ACM, pp. 287–294. doi:10.1145/130385.130417. 3

Data-Based DOIs	Description	Surveys & References
Clustering	**DOIs based on the results of clustering algorithms**	[Jai10]
Single Clustering	**DOIs based on the result of a single clustering algorithm**	[HBV02]
Cluster Characteristics	**DOIs based on characteristics of relations of instance to**	[HBV02, BZL*18]
Centroid Distance	Distance to (nearest/winning) cluster centroid	-
Cluster Crispness	Crispness: how clear an instance can be assigned to a single cluster	[HBV02]
Cluster Size Deviation	Difference of a cluster's size compared to the average cluster size	[STMT12]
Cluster Compactness	**DOIs based on within-cluster compactness (lower values are better)**	[HBV02]
Cluster Variance	Within-cluster variance / intra-cluster variance	[Dun74]
Dunn's Index Compact.	Dunn's Index: maximum within-cluster distance	[Rou87]
Silhuette Compactness	Silhuette Index: average within-cluster distance	[HBV02]
Cluster Separation	**DOIs based on between-cluster separation (higher values are better)**	[HBV02]
Other Centroids Distance	Accumulated distance to all other clusters	-
Dunn's Index Separation	Dunn's Index: minimum distance to nearest other cluster	[Dun74]
Silhuette Separation	Silhuette: Average distance to nearest other cluster	[Rou87]
Committee Results	**DOIs based on the results of multiple clustering algorithms**	
Centroid Distance	Accumulated distances to (nearest/winning) cluster centroids	-
Cluster Crispness	Accumulated crispness scores of multiple clustering results	-
Cluster Variance	Accumulated within-cluster variances / intra-cluster variances	-
Cluster Compactness	Accumulated cluster compactness scores of multiple clustering results	-
Cluster Separation	Accumulated cluster separation scores of multiple clustering results	-
Density	**DOIs based on the local data density in the vicinity of an instance**	-
kNN-Based	Accumulated similarity of k nearest neighbors	[BZL*18]
epsilon Neighbor Count	Number of neighbors in ε-region of an instance	[BZL*18]
epsilon Neighbor Distances	Relative distance to neighbors in ε-region of an instance	[BZL*18]
Spatial Balancing	Proximity of an instance to a set of given instances (training data, data coverage)	[BSB*15, BZL*18]
Outliers	**DOIs based on outlier detection**	[RRS00, CBK09]
kNN-Based	k nearest neighbors are used to assign outlier scores	[RRS00, BZL*18]
Outlier Analysis Model	Outlier score based on an upstream outlier analysis algorithm	[KN98, BKNS00, CBK09]

Table 1: *Data-based classes of degree-of-interest (DOI) functions. Inner branches of the taxonmy are encoded with bold font. Clustering-based, density-based, and outlier-based branches constitute the primary distinguishing characteristics for data-basd DOIs.*

[STMT12] SEDLMAIR M., TATU A., MUNZNER T., TORY M.: A taxonomy of visual cluster separation factors. *Computer Graphics Forum (CGF) 31*, 3pt4 (2012), 1335–1344. doi:10.1111/j.1467-8659.2012.03125.x. 2, 3

[VPS*02] VENDRIG J., PATRAS I., SNOEK C., WORRING M., DEN HARTOG J., RAAIJMAKERS S., VAN REST J., VAN LEEUWEN D. A.: Trec feature extraction by active learning. In *TREC* (2002). 3

[WKBD06] WU Y., KOZINTSEV I., BOUGUET J.-Y., DULONG C.: Sampling strategies for active learning in personal photo retrieval. In *IEEE International Conference on Multimedia and Expo* (2006), IEEE, pp. 529–532. doi:10.1109/ICME.2006.262442. 3

Model-Based DOIs	Description	Surveys & References
Uncertainty	**DOIs based on probability distributions for instances assigned by the classifier**	[Set12]
└Least Significant Confidence	High interestingness if probability of most confident class is low	[Set12]
└Smallest Margin	Score depending on the difference in probability between first two most confident classes	[WKBD06]
└Entropy	Score is based on the Entropy of the class distribution	[VPS*02]
Relevance	**DOIs based on the probability distributions for instances assigned by the classifier**	[Set12]
└Most Significant Confidence	High interestingness if probability of most confident class is high	[SC08]
Spatialization	**DOIs based on spatial information and relations between high-dimensional data**	
└**Class Relations**	**DOIs based on relations of instances to class characteristics (centroids, spread, etc.)**	
└**Class Characteristics**	**DOIs based on uncertainty caused by class spatialization**	
└Class Centroids Dist Margin	Smallest margin of distances to centroids of the winning and second most likely class	-
└Class Size Deviation	Difference of a class' size compared to the average class size (fosters balancing)	[STMT12]
└Class Borders	Likelihood of instances to be at the outbound of a class	[BZL*18]
└**Class Compactness**	**DOIs based on within-class compactness (lower values are better)**	
└Class Centroid Similarity	Distance of instances to the centroids of winning classes	-
└Dunn's Index Compactness	Dunn's Index: maximum within-class distance	[Dun74]
└Silhuette Compactness	Silhuette Index: average within-class distance	[Rou87]
└**Class Separation**	**DOIs based on between-class separation (higher values are better)**	
└Class Centroids Distances	Probability-weighted distances to centers of non-winning classes	-
└Dunn's Index Separation	Dunn's Index: minimum distance to nearest other class	[Dun74]
└Silhuette Separation	Silhuette Index: average distance to nearest other class	[Rou87]
└**Neighbor Relations**	**DOIs based on neighbor instances**	
└**Neighbor Votes**	**DOIs based on the diversity of winning class labels (votes) of k nearest neighbors**	
└Vote Cardinality	Number of different votes among the k nearest neighbors	-
└Vote Entropy	Entropy of votes	[Sha48]
└Simpson Diversity	Simpson's Diversity index of votes	[Sim49]
└Winner Vote Count	Number of votes of the most voted class	-
└**Neighbor Probabilities**	**DOIs based on the comparison of probability distributions among k-NN**	
└Probability Distance	Euclidean distance to neighbors' probability distributions	-
└Kullback Leibler Div.	Kullback-Leibler divergence of neighbors' probability distributions	[KL*51]
└Jensen Shannon Divergence	Jenson-Shannon divergence neighbors' probability distributions	[FT04]
└Kolmogorov Smirnov Dist.	Kolmogorov-Smirnov test neighbors' probability distributions	[Kol33, Smi48]
└**Neighbor Prob. Aggregation**	**DOIs based on aggregated probability distributions among k-NN**	
└Least Significant Confid.	High interestingness if probability of most confident class is low	-
└Smallest Margin	Score depending on the difference in probability between first two most confident classes	-
└Entropy	Score is based on the Entropy of the class distribution	-
Committees	**DOIs based on a committee of classification models**	[SOS92, Set12]
└**Votes**	**DOIs based on the diversity of winning class labels (votes) of the committee**	[SOS92, Mam98]
└Vote Cardinality	Number of different votes among the k nearest neighbors	-
└Vote Entropy	Entropy of votes	[Sha48]
└Simpson Diversity	Simpson's Diversity index of votes	[Sim49]
└**Probabilities**	**DOIs based on the divergence of probability distributions proposed by the committee**	[Set12]
└Probability Distance	Euclidean distance to neighbors' probability distributions	-
└Kullback Leibler Divergence	Kullback-Leibler divergence of neighbors' probability distributions	[KL*51]
└Jensen Shannon Divergence	Jenson-Shannon divergence neighbors' probability distributions	
└Kolmogorov Smirnov Dist.	Kolmogorov-Smirnov test neighbors' probability distributions	[Kol33, Smi48]

Table 2: *Model-based classes of degree-of-interest (DOI) functions. Inner branches of the taxonmy are encoded with bold font. Uncertainty-based, relevance-based, spatialization-based, and committee-based branches constitute the primary distinguishing characteristics for model-basd DOIs.*