2.3.1 Sufficiency of training data

Practical guidance – cross-domain

Authors: Rob Ashmore (Dstl), and Dr Radu Calinescu and Dr Colin Paterson (Assuring Autonomy International Programme)

Fundamentally, training an ML model starts with data. These data describe the desired relationship between the ML model inputs and outputs, the latter of which may be implicit for unsupervised approaches. Equivalently, these data encode the requirements we wish to be embodied in our ML model. Consequently, any assurance argument needs to explicitly consider data.

Although the objective here is to produce data sets for training the model, the guidance provided here is also applicable to generating verification data to be used as part of the model verification stage (Guidance on model verification is provided in section 2.3.3).

Stage Input and Output Artefacts

The key input artefact to the Data Management stage is the set of requirements that the model is required to satisfy. These may be informed by verification artefacts produced by earlier iterations of the ML lifecycle. The key output artefacts from this stage are data sets: there is a combined data set that is used by the development team for training and validating the model; there is also a separate verification data set, which can be used by an independent verification team.

Activities

1. **Collection** - This data-management activity is concerned with collecting data from an originating source. These data may be subsequently enhanced by other activities within the Data Management stage. New data may be collected, or a pre-existing data set may be re-used (or extended). Data may be obtained from a controlled process, or they may arise from observations of an uncontrolled process: this process may occur in the real world, or it may occur in a synthetic environment.
2. **Preprocessing** – It is assumed here that preprocessing is a one-to-one mapping: it adjusts each collected (raw) sample in an appropriate manner. It is often concerned with standardising the data in some way, e.g. ensuring all images are of the same size [1]. Manual addition of labels to collected samples is another form of preprocessing.
3. **Augmentation** - Augmentation increases the number of samples in a data set. Typically, new samples are derived from existing samples, so augmentation is, generally, a one-to-many mapping. Augmentation is often used due to the difficulty of collecting observational data (e.g. for reasons of cost or ethics [2]). Augmentation can also be used to help instil certain properties in the trained model, e.g. robustness to adversarial examples [3].
4. **Analysis** - Analysis may be required to guide aspects of collection and augmentation (e.g. to ensure there is an appropriate class balance within the data set). More

generally, exploratory analysis is needed to provide assurance that Data Management artefacts exhibit the desired properties discussed below.

Desired Assurance Properties

From an assurance perspective, the data sets produced during the Data Management stage should exhibit the following key properties:

1. **Relevant** - This property considers the intersection between the data set and the desired behaviour in the intended operational domain. For example, a data set that only included German road signs would not be Relevant for a system intended to operate on UK roads.
2. **Complete** - This property considers the way samples are distributed across the input domain and subspaces of it. In particular, it considers whether suitable distributions and combinations of features are present. For example, an image data set that displayed an inappropriate correlation between image background and type of animal would not be complete [4].
3. **Balanced** - This property considers the distribution of features that are included in the data set. For classification problems, a key consideration is the balance between the number of samples in each class [5]. This property takes an internal perspective; it focuses on the data set as an abstract entity. In contrast, the Complete property takes an external perspective; it considers the data set within the intended operational domain.
4. **Accurate** - This property considers how measurement (and measurement-like) issues can affect the way that samples reflect the intended operational domain. It covers aspects like sensor accuracy and labelling errors [6]. The correctness of data collection and preprocessing software is also relevant to this property, as is configuration management.

Methods

Table 1 provides a summary of the methods that can be applied during each Data Management activity in order to achieve the desired assurance properties (desiderata). Further details on the methods listed in Table 1 are available in [7].

Method	Associated activities†				Supported desiderata‡			
	Collection	Preprocess.	Augment.	Analysis	Relevant	Complete	Balanced	Accurate
Use trusted data sources, with data-transit integrity guarantees	✔				★			
Experimental design [8, 9]	✔		✔		★	★	☆	
Simulation verification and validation [10]			✔		★	☆	☆	
Exploratory data analysis [11]				✔		★	★	
Use adversarial examples [12]			✔		☆	★		
Include a "dustbin" class [13]			✔		☆	★		
Remove unwanted bias [14]		✔	✔		★		☆	
Compare sampling density [15]			✓	✔		★	☆	
Identify empty and single-class regions [16], [17]			✓	✔		★	☆	
Use situation coverage [18]				✔		★		
Examine system failure cases				✔		★		
Oversampling & undersampling [19]				✔		★	★	
Check for within-class [20] and feature imbalance				✔		★		
Use a GAN [21]			✔			★	☆	
Augment data to account for sensor errors	✓		✔		☆			★
Confirm correct software behaviour [22, 23]	✓	✔	✔	✓	☆	★	☆	☆
Use documented processes	✔	✔	✔	✔	☆			★
Apply configuration management [22, 23]	✔	✔	✔	✔	☆			★

†✔ = activity that the method is typically used in; ✓ = activity that may use the method
‡★ = desideratum supported by the method; ☆ = desideratum partly supported by the method

Table 1 – Assurance methods for data management

Summary of Approach

1. Take the requirements that the model is required to satisfy.
2. Apply appropriate methods in order to undertake each data management activity to generate training data that achieves the desired assurance properties whilst satisfying the requirements.
    a. Apply appropriate methods for data collection
    b. Apply appropriate methods for preprocessing data
    c. Apply appropriate methods for augmentation of data
    d. Apply appropriate methods for data analysis
3. Apply appropriate methods in order to undertake each data management activity to independently generate verification data that achieves the desired assurance properties whilst satisfying the requirements (Guidance on the selection of verification data is provided in section 2.3.3)

References

- [1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. IEEE 86, 11 (1998), 2278–2324.
- [2] German Ros, Laura Sellart, Joanna Materzynska, et al. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In IEEE Conf. on computer vision and pattern recognition. 3234–3243.
- [3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. (2014). arXiv:1412.6572

- [4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining. ACM, 1135–1144.
- [5] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73 (2017), 220–239.
- [6] Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. Journal of Artificial Intelligence Research 11 (1999), 131–167.
- [7] Ashmore, R., Calinescu, R. and Paterson, C., 2019. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. arXiv preprint arXiv:1905.04223.
- [8] Roger E Kirk. 2007. Experimental design. The Blackwell Encyclopedia of Sociology (2007).
- [9] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. 1989. Design and analysis of computer experiments. Statistical science (1989), 409–423.
- [10] Robert G Sargent. 2009. Verification and validation of simulation models. In Winter Simulation Conf. 162–176.
- [11] John W Tukey. 1977. Exploratory data analysis. Vol. 2. Reading, Mass.
- [12] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, et al. 2017. Practical black-box attacks against machine learning. In Asia Conf. on Computer and Communications Security. ACM, 506–519.
- [13] Mahdieh Abbasi, Arezoo Rajabi, Azadeh Sadat Mozafari, Rakesh B Bobba, and Christian Gagne. 2018. Controlling Over-generalization and its Effect on Adversarial Examples Generation and Detection. (2018). arXiv:1808.08282
- [14] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, et al. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. (2018). arXiv:1810.01943
- [15] Arijit Bishnu, Sameer Desai, Arijit Ghosh, Mayank Goswami, and Paul Subhabrata. 2015. Uniformity of Point Samples in Metric Spaces Using Gap Ratio. In 12th Annual Conf. on Theory and Applications of Models of Computation. 347–358.
- [16] Joseph Lemley, Filip Jagodzinski, and Razvan Andonie. 2016. Big holes in big data: A Monte Carlo algorithm for detecting large hyper-rectangles in high dimensional data. In IEEE Computer Software and Applications Conf. 563–571.
- [17] Rob Ashmore and Matthew Hill. 2018. Boxing Clever: Practical Techniques for Gaining Insights into Training Data and Monitoring Distribution Shift. In Int. Conf. on Computer Safety, Reliability, and Security. Springer, 393–405.
- [18] Rob Alexander, Heather Rebecca Hawkins, and Andrew John Rae. 2015. Situation coverage—a coverage criterion for testing autonomous robots. Technical Report YCS-2015-496. Department of Computer Science, University of York.
- [19] Victoria López, Alberto Fernández, Salvador García, et al. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250 (2013), 113–141.
- [20] Nathalie Japkowicz. 2001. Concept-learning in the presence of between-class and within-class imbalances. In Conf. of the Canadian Society for Computational Studies of Intelligence. Springer, 67–77.
- [21] Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. (2017). arXiv:1711.04340

- [22] ISO. 2018. Road Vehicles - Functional Safety: Part 6. Technical Report BS ISO 26262-6:2018. ISO.
- [23] RTCA. 2011. Software Considerations in Airborne Systems and Equipment Certification. Technical Report DO-178C.