International Journal of Machine Learning and Cybernetics (2023) 14:1–2
https://doi.org/10.1007/s13042-022-01699-0

EDITORIAL

Special issue on small data analytics

Hui Wang[1] · Ivo Duentsch[2] · Gongde Guo[3] · Sadiq Ali Khan[4]

Published online: 27 November 2022

Big data is an important characteristic of the modern time. Very large data sets can be collected at reasonable costs – for example, they can be captured from sensors, which are widely available, or crowd sourced from the public. However, large data sets with annotation (labelling, structuring etc.) are very rare, as the annotation process can be very expensive. This is especially the case in biomedical domains, where proper annotation must be done by experts. Therefore, data sets with annotation are mostly small. This poses special challenges in data analysis and addressing such small data challenges needs special methodologies in what can be called *small data analytics*. There are already some specialist research areas that are related to small data analytics, such as one-shot, zero-shot or few-shot learning, manifold learning, imbalanced learning, transfer learning, active learning, model-based learning, data augmentation, regularisation and visualisation.

Three technical challenges of small data analytics can be identified. The first challenge is *to effectively learn with small data sets*. Machine learning, especially deep learning, can effectively learn with large data sets. However, it cannot effectively learn with small data sets due to various issues such as overfitting, noise, outliers and sampling bias, which can render the learned model ineffective. There are ways to deal with small data sets, such as data augmentation, transfer learning, regularisation and visualisation. However, these methods need skilled people to use, and their effectiveness is limited. The second challenge is *to annotate data more easily and cheaply* so that large data sets with annotation are not scarce assets anymore. The annotation challenge can be addressed via the use of annotation tools/services in a number of ways. One way is to provide annotation tools such as Computer Vision Annotation Tool so that annotation can be done more effectively and easily; another is to outsource the annotation task to an annotation service provider such as Amazon Mechanical Turk. However, none of these approaches actually solves the annotation problem scientifically; rather they do so as a business. To solve the annotation problem scientifically, one desirable approach is to design a new machine learning algorithm that needs minimal feedback from human experts. This will only be possible if domain specific constraints (i.e. domain knowledge) can be imposed in the learning process to reduce the model space hence the variance in learning. The third challenge is *to leverage domain knowledge (common sense, domain specific) implicitly or explicitly in the learning process*. When there is a large amount of data, machine learning needs a small amount of knowledge (e.g. in choosing a machine learning model); when there is not much data, machine learning needs a large amount of knowledge to reduce the model space. It is therefore reasonable to hypothesise that knowledge-based machine learning is a way forward for small data analytics.

This Special Issue publishes 13 papers on small data analytics which address the first and third technical challenges identified above by data augmentation, metric learning, Bayesian modelling and using domain knowledge. It also features contemporary applications where small data is a challenge or where both knowledge and data are needed.

In "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers", Bayer *et al*. present a study on data augmentation for text classification using small datasets. They propose a text generation method for data augmentation. Text classification results on small datasets using the proposed text generation method are quite promising, achieving up to 15.53% increase from the baseline without augmentation. They discuss the generality and limitation of the method as well as future work.

In "Tuning of data augmentation hyperparameters in deep learning to building construction image classification

✉ Hui Wang
h.wang@qub.ac.uk

[1] Queen's University, Belfast, UK

[2] Brock University, St. Catharines, Canada

[3] Fujian Normal University, Fuzhou, China

[4] University of Karachi, Karachi, Pakistan

with small datasets", Ottoni *et al*. study data augmentation for building construction image classification using small data. They propose a method for tuning data augmentation hyperparameters. In their experimental evaluation, they consider 128 combinations of transformation in the image, and achieve up to 95.6% in the image classification task using the hyperparameters recommended by the proposed method.

In "Data augmentation for aspect-based sentiment analysis", Li *et al*. present a study of data augmentation for aspect-based sentiment analysis (ABSA). They propose two augmentation strategies: part-of-speech (PoS) wise synonym substitution (PWSS) and dependency relation-based word swap (DRAWS), which augment data by using PoS, external domain knowledge, and syntactic dependency. Their experiments on small data sets show that these strategies achieved up to 11.49% increase in macro-F1 over the baseline without data augmentation.

In "Distance metric learning with local multiple kernel embedding", Tsang *et al*. address the small data problem via distance metric learning. They propose a new metric learning method for small datasets, *distance metric learning with local multiple kernel embedding* (DMLLMK), which employs multiple kernel functions. Instead of weighting multiple kernels equally, which is common in the literature, they propose the weight of each kernel function in DMLLMK be assigned locally based on the input data. This local weight method enables metric learning to capture more information in the data. The metric learning and local weighting are optimized simultaneously by an alternating learning process. The DMLLMK is applicable to small datasets via constraining and regularization to keep DMLLMK conservative and prevent it from overfitting on small data. Their experimental results show the proposed method is effective and can outperform other metric learning methods on small datasets.

In "Merits of Bayesian networks in overcoming small data challenges: a meta-model for handling missing data", Ameur *et al*. address the small data problem through modelling small data with Bayesian networks, in case the data has missing values. They propose a new strategy for learning multiple Bayesian network structures and a novel method for the weighted fusion of Bayesian network structures. They introduce four systems to address separately the missing values/records involving annotation, balancing, missing values imputation and data over-sampling. They further combine these systems into a meta-model. Experimental results demonstrate the capability of the proposed method to deal with multi-class problems and with extremely small datasets.

In "Using a small dataset to classify strength-interactions with an elastic display: a case study for the screening of autism spectrum disorder", Tentori *et al*. present a case study using a small dataset and domain-specific knowledge for the screening of Autism Spectrum Disorder (ASD). Health data collection of children with ASD is challenging, time-consuming, and expensive; thus, working with small datasets is inevitable in this area. The diagnosis rate in ASD is low, leading to several challenges, including imbalance classes, potential overfitting, and sampling bias, making it difficult to show its potential in real-life situations. The authors collected data from 68 children, extracted features and selected the most relevant ones. They found the proposed machine learning models can correctly classify children with ASD with 97.3% precision and recall even if the classes are unbalanced. Increasing the size of the dataset via synthetic data improved the model precision to 99%. They finish discussing the importance of leveraging domain-specific knowledge in the learning process to successfully cope with some of the challenges faced when working with small datasets in a concrete, real-life scenario.

In "Fractional mega trend diffusion function-based feature extraction for plant disease prediction", Bhatia *et al*. address the small data problem from feature extraction point of view in a plant disease early detection task. They propose a novel feature extraction technique that can produce more relevant features for small-sized datasets by performing some operations on the original features. Two small plant diseases datasets are used to evaluate the proposed approach. Resampling techniques are used to balance the imbalanced datasets and Optimized Kernel Extreme Learning Machine (OKELM) algorithm is used for the classification. The results of their experiment show that the proposed approach achieved between 70% and 89.47% in accuracy for one dataset and between 92.27% and 100% for the other. They conclude the proposed approach performed well for the small datasets.

In summary, the papers in this special issue address the small data problem from an applied point of view. The editors hope that the papers in this special issue will lead to further research in the practical as well the theoretical aspects of handling small data.