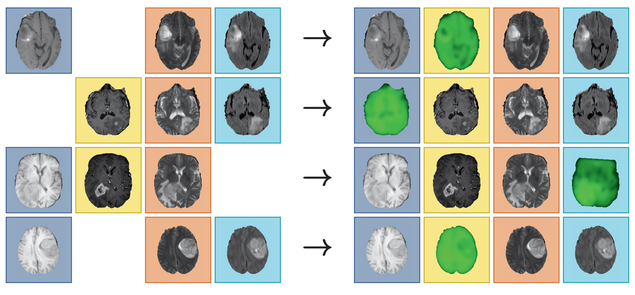
Why does synthesized data improve multi-sequence classification?

Gijs van Tulder and Marleen de Bruijne

Motivation

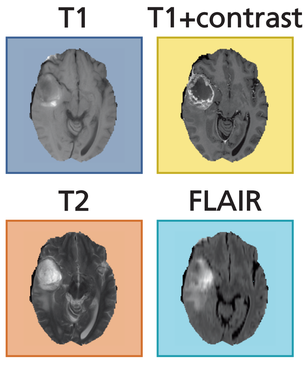
Data synthesis methods can be used to predict missing modalities in incomplete datasets. Some people have found that using synthetic data can improve classification accuracy.

But synthesis models mostly transform existing information. Why does data synthesis still help classification?



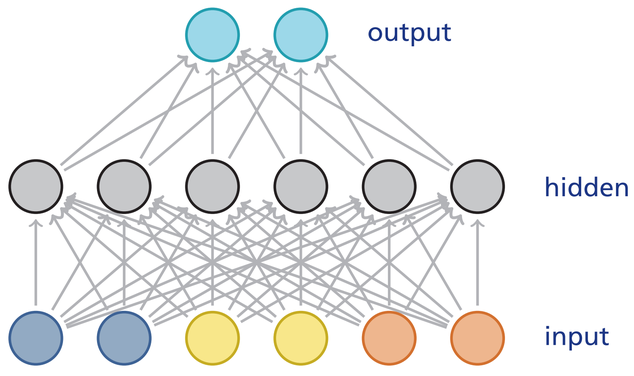
Experiments

We did experiments with two synthesis methods and two classifiers.
We used data from the BRATS 2013 brain tumor segmentation dataset.



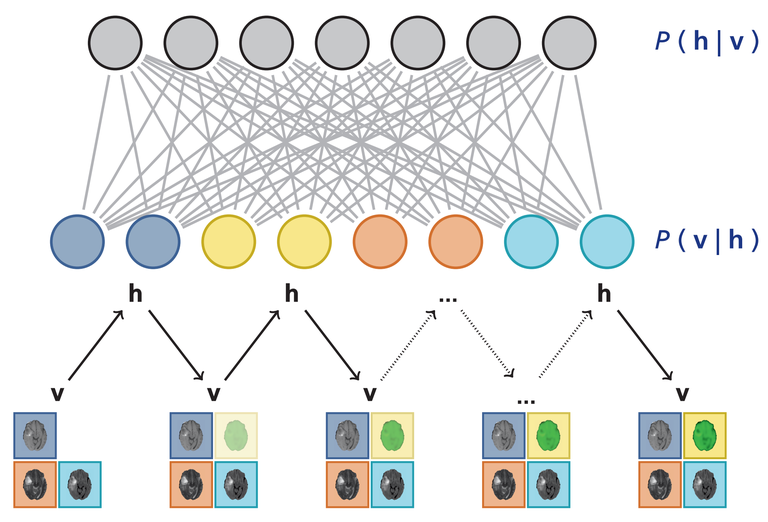
Method 1: Neural networks

Neural networks can be trained to predict one modality from the others. Because a network predicts only one modality, you need multiple networks to predict all modalities.



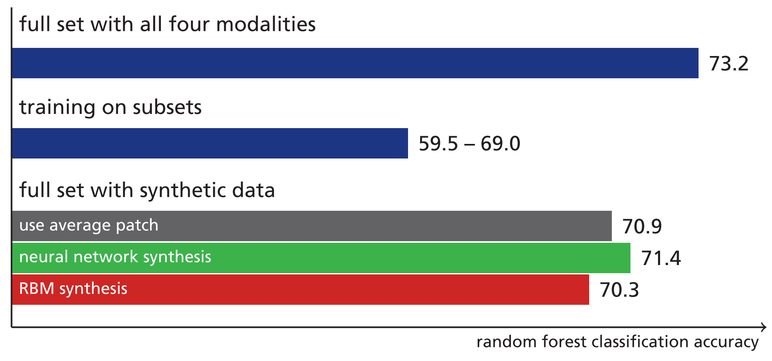
Method 2: Restricted Boltzmann machines (RBMs)

With RBMs, a single model can be used to predict any of the sequences.



Explanation 1: Synthesis allows more training data

Using synthetic data, you can train a single classifier on samples with different missing modalities. This gave better results than training on subsets.



Conclusions

Synthetic data can improve classification accuracy when training with incomplete datasets.

We proposed synthesis using RBMs to naturally handle incomplete samples.

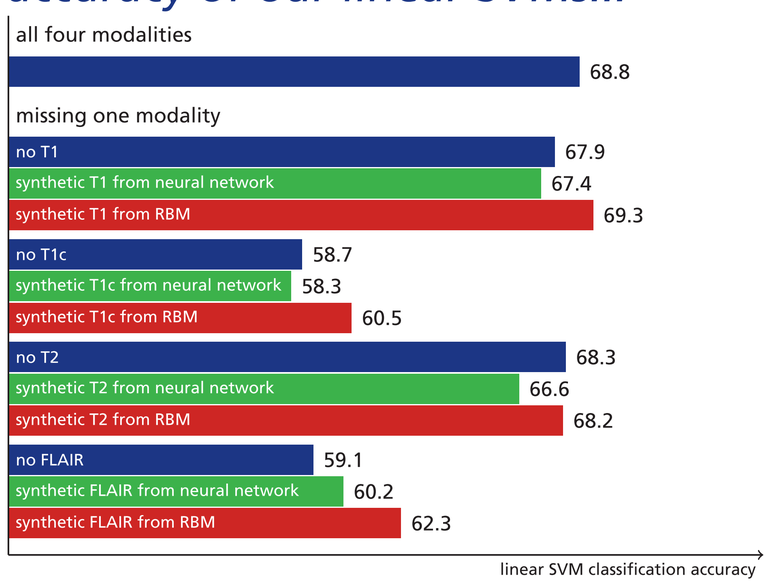
Synthetic data can help because it allows training with more samples and can provide useful transformations.

This depends on the type of classifier.

Explanation 2: Synthesis models provide new transformations of the data

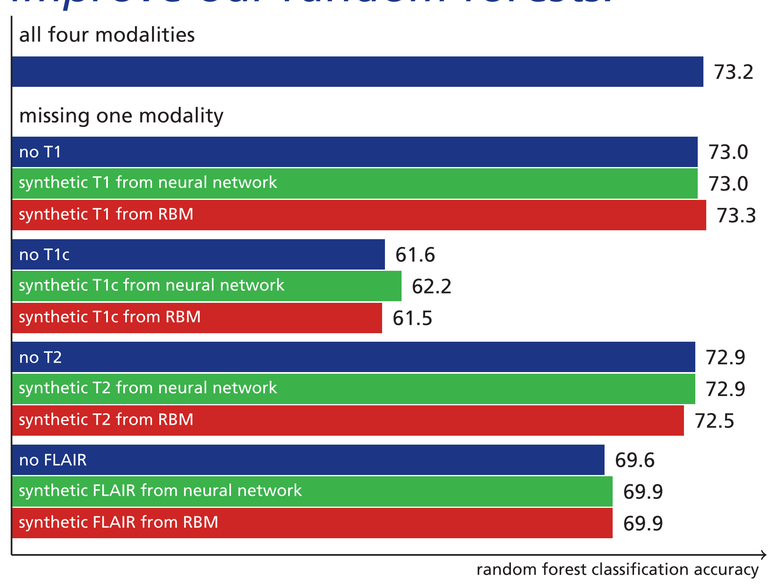
Data synthesis models can provide non-linear transformations of the data, which can be useful for classifiers that cannot do this on their own.

Synthetic data improved the accuracy of our linear SVMs...



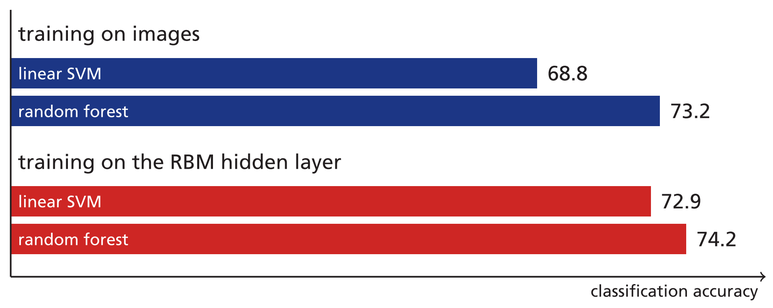
More flexible classifiers can extract the same information from the original data and will not become much better from using synthetic data.

but synthetic data did not improve our random forests.



The classifiers can also be trained with values from the RBM's hidden layer. This is a non-linear transformation of the input, so it will be more useful to a simple linear classifier than to more flexible classifiers.

Training on the RBM's hidden layer helped the linear SVMs more than the random forests.



Biomedical Imaging Group Rotterdam, the Netherlands
www.bigr.nl · g.vantulder@erasmusmc.nl

MICCAI
October 2015, München, Germany

This research is financed by the Netherlands Organization for Scientific Research (NWO).

Erasmus MC
University Medical Center Rotterdam