Big Data

Eliciting reliable information from big data with classifiers and multimodal data fusion

Prof. Hans A. Kestler knows a great deal about large amounts of data. He heads up the Institute of Medical Systems Biology at the University of Ulm and is constantly inundated with cooperation enquiries from clinicians. On behalf of BIOPRO, Walter Pytlik asked him whether the conditions for using big data more in biomedical research are already largely present.

Prof. Hans A. Kestler
© University of Ulm

What do you feel are the biggest challenges we face as we move towards integrated healthcare systems?

Getting there is a process of continuous improvement. I think that the biggest challenges are to do with data protection, the many legal aspects such as data ownership, as well as technical issues. With regard to semantics, the biggest challenge is to take data from different locations, and even from within the same hospital, and make them comparable.

Let's start with data protection.

I'm not an expert in this area. But I see problems, for example, as far as the transfer of data between different countries is concerned, because this involves many stakeholders – cost bearers, clinics and data protection officers. This could lead to concerns about data integration, and not just for data privacy reasons.

What do you mean?

Is data integration really what we want when you consider the eventuality that this will allow people to draw conclusions about individual clinics or departments? Take, for example, the integration of treatment data. This could potentially bring to light different types of treatment used in different locations. It could then become apparent that some of the treatment methods used do not follow valid guidelines. This could lead to different treatment methods being compared erroneously, when in fact they are not comparable at all due to differently structured patient groups in individual clinics.

If combining different data enables conclusions to be drawn on treatment performance, treatment collectives and other performance-relevant parameters, data integration might become difficult to implement. Legal measures would probably need to be put in place, and this would only be possible by legally requiring clinics and cost bearers to provide data under certain conditions. The problem is the details. If implementation is not solved on the detail level and stakeholders do not pull together, then any attempt to integrate data will effectively be blocked.

Nonetheless, data integration will come sooner or later; it has to. It is too big an opportunity to miss in terms of improving therapies. It can help identify new molecular subgroups. Once patient data from more than just one clinic becomes available, treatment is likely to be more successful. It is a key issue in the case of rare diseases where data integration is already the order of the day.

Does the patient own this data?

Yes - but it is not currently a huge concern for many researchers because they are often working with anonymised and pseudonymised data. This will change as more intelligent methods such as artificial intelligence and decision support enter clinical settings. In fact, methods like deep learning, artificial neural networks and other machine learning methods are already being used.

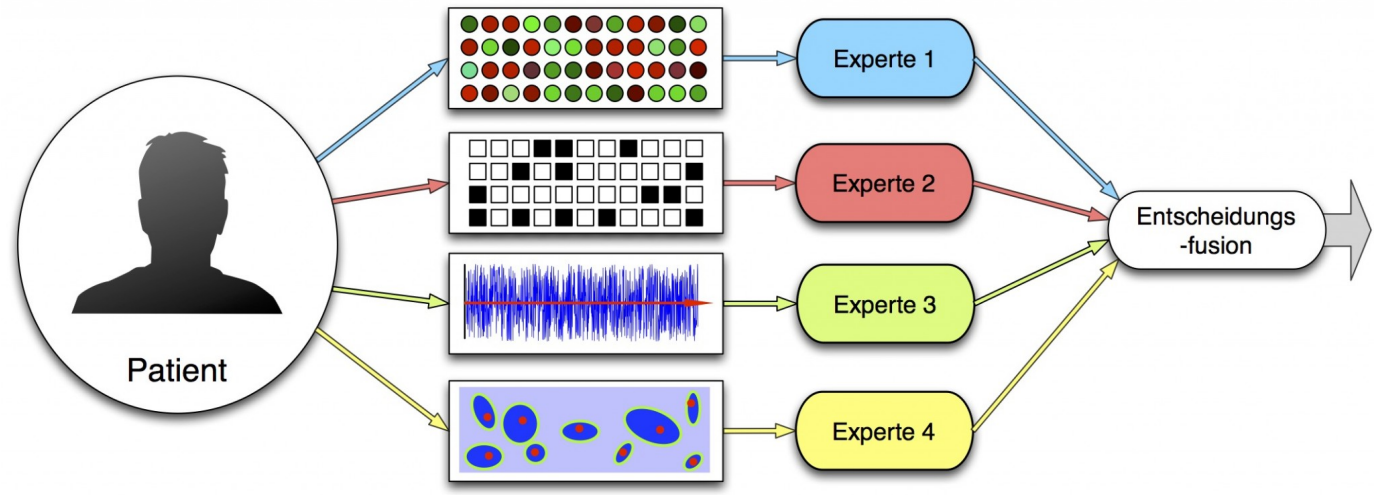
What needs to be done with the data in order to create meaningful results using the aforementioned methods? How do big data become smart data?

My group and I are very interested in interpretable decision-making processes and are looking for ways to generate these processes automatically. This is called interpretable classification. We do not want the black box problem often associated with deep learning. We want to build simple, well-generalising classifiers from whose structure we can learn, so that we can, for example, find out which molecular pathway is important for classification or which clinical data are key for making a treatment decision. In short, we are building classifiers that can generalise as effectively as possible to new data.

How does this work?

This is how this kind of classification process works: we practise with existing data, adapt a system and apply this system to new data. For example, we want to use the classification of one tumour entity to draw conclusions on another tumour entity. Classifiers learning from each other is, in turn, a sub-topic of research into this area.

In our classification approaches, we are modelling tumour boards to a certain extent. These are interdisciplinary tumour conferences held in comprehensive oncology centres. We have a range of different algorithm experts. One expert looks at the data from one point of view only, the next one from yet a different one. They then make a consensus decision. We formulate such procedures algorithmically. This is what our research is about. If you know that only one particular subset of experts is required to make a particular decision, conclusions can be drawn about the genesis of



Schematic showing multimodal data fusion that can be used to combine knowledge from different abstraction levels with each other. The goal is to reduce the complexity of the classifier and thus improve its generalisation performance.

the disease.

Another approach is multimodal data fusion. Data from different modalities (see Figure) is combined. This allows knowledge from different levels of abstraction to be linked. The challenge is the different levels of knowledge. Example: A gene is part of one or more signalling pathways. This information can help us reduce the complexity of an RNA sequencing classifier and thereby improve its generalisation performance. That's what we do with multimodal data fusion.

We hope that this will reduce complexity and lead to better generalisability. In the end, it always comes down to developing methods that can make effective generalisations, i.e. are able to respond as well as possible to new data, even if we have never used this data for training purposes.

Is there a gold standard of medical documentation, from laboratory values to sequencing data?

No, there are many standards, not just one. This is both good and bad. As welcome as such a gold standard could be, it would also be restrictive. This is because the field of technology is not static, it is in constant flow. If you were to immobilise many things, effective progress might not necessarily be possible. Even if there were a newer, better technology, you would have to use the old one as stipulated by the relevant standard.

There have been various attempts at standardisation, including in the area of laboratory diagnostics. However, not everybody applies these standards. And there are always reasons not to apply them. This will continue to be a challenge. Minimal values, such as data on insurance cards, are easy to combine. But as far as blood values are concerned, the diagnostic methods used for obtaining these values will most likely differ. So combining this type of values will be a greater challenge.

Does this require linking data with as much concise and exact metadata as possible?

Many metadata are already available. The quantity very much depends on the issue under investigation. Sequence data, for example, can be linked with additional information via widely available databases such as KEGG*. This kind of general knowledge is constantly changing and helps to structure and group data. We can even try to integrate this abstract knowledge into

classification processes. This has led to very good results in our research.

So you provide the data structure, ontologies as you would probably say...?

Yes, ontologies are another option. This can go even further, for example in cases when details about signalling networks are known. Such data contain a lot of structural information. How many of these structures are suitable for inclusion in classification processes is something that ongoing research projects are investigating.

Does this mean that there are already tools and methods that can be used to extract meaningful information from data, i.e. producing correlations?

The ontologies and pathway examples I have given are not correlations. Here, genes are associated with a certain terminology, which in turn can help to pre-structure and better classify the data. This pre-structuring is intrinsic to image data, where structures can be discerned easily: lines, shades of grey, etc. As far as genomic data is concerned, this structure can be deduced from the chromosomal location of a gene, but not necessarily directly from its functionality. That said, there are lists of genes that can be placed in functional groups, oncogenes, for example. One therefore tries to use data structures in the classification process. Another aspect is the integration of image data. There are numerous methods, many of which are being developed further. As far as algorithms are concerned, I think we've already come quite a long way, although they are not yet comprehensively applied.

So you would first have to solve the problems mentioned above? How can data be merged, exchanged and shared?

Exactly. How can they be anonymised and pseudonymised, so that they can be used effectively? I think that the way that access to this data is managed plays a key role in the processes. Algorithms are developed by computer scientists. This is something that they like doing, and that they can do relatively quickly. We are talking big data, especially as far as genomic data is concerned. This is nothing like data that originates from technical processes.

* Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/)

Article

26-Mar-2018
Walter Pytlik

Further information

Prof. Dr. Hans A. Kestler
Director Institute for Medical Systems Biology
Ulm University
Phone: +49 (0)731 50024500

- Medical Systems Biology

The article is part of the following dossiers

Data mining: new opportunities for medicine and public health

Big data - the big promise of the new digitised world

bioinformatics | database | telemedicine | data mining | big data