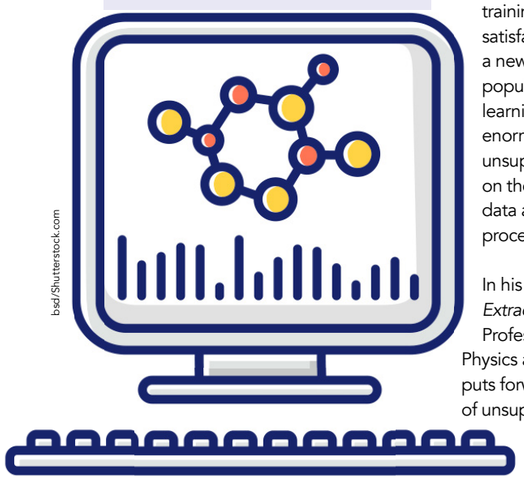
Unsupervised feature extraction applied to bioinformatics

In his new book, Professor Y-h Taguchi, from Chuo University, Tokyo, Japan, takes two classical mathematical techniques, principal component analysis and tensor decomposition, and demonstrates how they can be used to perform feature selection in his cutting-edge research. Both unsupervised learning methods are applied to carry out feature extraction in a wide range of 'large p small n' problems. This book enables the reader to analyse data sets with small samples and many features. It offers a fast algorithm for analysing big data with output that is easily interpreted.



bsd/Shutterstock.com

Machine learning, an artificial intelligence (AI) application that automates analytical model building for data analysis, has gained recent popularity as it can handle massive datasets. It is based on the concept that systems can learn, identify patterns and make decisions from data with little or no human intervention.

Machine learning techniques are powerful devices that benefit a wide spectrum of applications such as bioinformatics, where they are employed to explore the underlying mechanisms and interactions between biological molecules. They can also be used to predict the structure and function of proteins and identify disease-causing mutations. Moreover, they are essential tools in the biomarker discovery process.

Supervised learning techniques, such as deep learning, learn the answers from a training dataset and have to achieve a satisfactory performance level before a new dataset is introduced. While popular, the application of supervised learning is limited as it requires an enormous amount of data. In contrast, unsupervised learning methods work on their own to discover patterns in the data and can perform more complex processing tasks.

In his new book, *Unsupervised Feature Extraction Applied to Bioinformatics*, Professor Y-h Taguchi, Professor of Physics at Chuo University, Tokyo, Japan, puts forward his novel applications of unsupervised learning. Professor Taguchi takes two classical mathematical techniques, principal component analysis and tensor decomposition, and demonstrates how they can be used to perform feature selection in his cutting-edge research.

FEATURE SELECTION

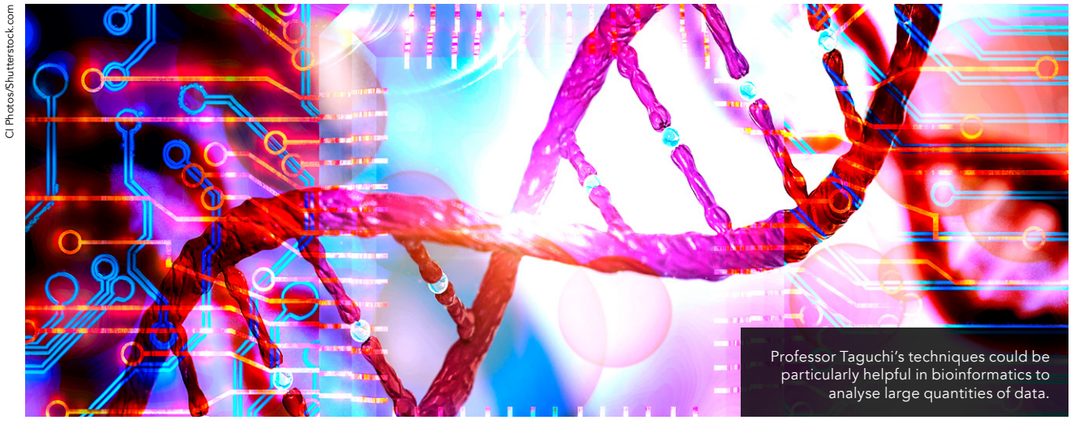
Feature selection is the process of selecting a subset of relevant features, such as variables, attributes or predictors, from a dataset. Feature selection methods are frequently employed when there are a large number of features and relatively few samples to analyse. For example, it is often used in the analysis of DNA microarray data where just tens or hundreds of samples contain many thousands of features.

THE CURSE OF DIMENSIONALITY

Professor Taguchi is particularly interested in feature selection involving the selection of a small number of critical variables from a huge number of variables. These problems are known as 'large p small n' problems. 'Large p' refers to the massive number of variables and 'small n' denotes the very small number of available observations. This predicament is known as 'the curse of dimensionality' which refers to the lack of enough observations when compared with the dimensionality, i.e., the number of relevant features. This is a problem that has remained unresolved for a long time.

A JOURNEY FOR THE READER

Professor Taguchi has designed his book so that the intended audience, including students, researchers, and practitioners, can grasp his concepts with ease. He has organised the book into three sections, taking the reader on a journey from the rudimentary mathematical foundations through feature extraction techniques and on to the application of his concepts to various data problems in genomic science.



CJ Photos/Shutterstock.com

Professor Taguchi's techniques could be particularly helpful in bioinformatics to analyse large quantities of data.

MATHEMATICAL PREPARATIONS

The book starts with Professor Taguchi reintroducing basic concepts of linear algebra including scalars, vectors, matrices, and tensors, written from data science points of view. He describes how vectors are sets of scalars, or real numerical values, matrices are composed of vectors and tensors are made up of matrices and goes on to demonstrate their arithmetic manipulation. The generation of new features as well as the use of dummy scalars and vectors to represent descriptive features such as colours that cannot be defined with real values are covered too. These early chapters also contain exercises with worked solutions so the reader can practice the methods and check their understanding.

> ***PCA and TD-based unsupervised feature extraction methods are powerful tools in the study of biological problems involving biomarker identification, gene expression, and drug discovery.***

The reader is then introduced to principal component analysis (PCA), as a part of matrix factorisation, and tensor decomposition (TD), two embedding techniques that can reduce dataset dimensionality.

FEATURE EXTRACTION

Professor Taguchi introduces feature extraction, a data-driven generator of new features. Feature extraction reduces the number of features in a dataset by creating new features from combinations of the original features, which are then discarded. The new reduced set of features should summarise the majority of the original information. This differs from feature selection, which tends to rank the importance of the original features and discard the less important, redundant and irrelevant ones. (Feature selection doesn't create new features.)

The author explains how principal component analysis and tensor decomposition can be used in various feature extraction and feature selection procedures and demonstrates their application to synthetic datasets.

FEATURE EXTRACTION VS FEATURE SELECTION

The choice of whether to use feature extraction or feature selection is complex and depends highly on the context of the problem rather than mathematical considerations. Professor Taguchi's targeted field is bioinformatics and he explains the reasons behind the preference for feature selection. In bioinformatics analysis there is often a cost associated with measuring individual features, so measuring fewer features can reduce the cost of individual observations. This means that the number of observations can be increased, which can lead to a better outcome.

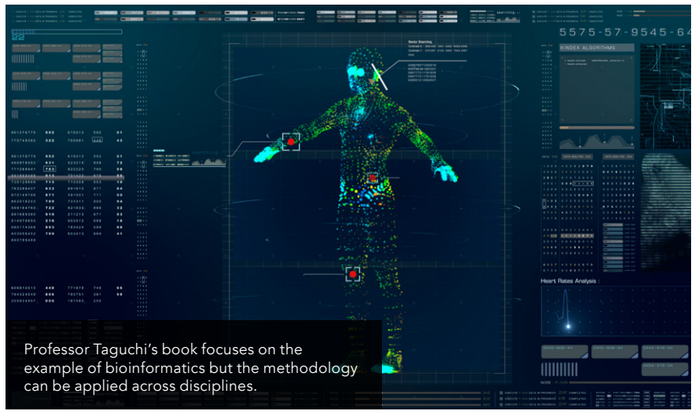
Even when cost is not an issue, feature selection is often preferable because each feature maintains its own meaning. The author uses an example where genes are features, and notes that it is more useful to have a limited number of genes that can be interpreted rather than features combining a large number of genes that cannot be explained.

APPLICATIONS TO BIOINFORMATICS

In the final section of his book, Professor Taguchi provides an introduction to genomic science before demonstrating the application of two methods, principal components analysis (PCA)-based unsupervised feature extraction and tensor decomposition (TD)-based unsupervised feature extraction, to large p small n problems in bioinformatics.

Bioinformatics focuses on the study of DNA, mRNA (messenger ribonucleic acid or RNA) and proteins. Much of the analysis in the latter chapters of this book concentrates on DNA methylation and miRNA expression data.

DNA methylation is a biological process involving the addition of methyl groups to DNA molecules that is essential for



Color4260/Shutterstock.com

Professor Taguchi's book focuses on the example of bioinformatics but the methodology can be applied across disciplines.

The methodology, based upon linear algebra, offers a fast algorithm for analysing big data with output that is easily interpreted.

normal development. DNA methylation is associated with fundamental processes such as genomic imprinting, aging, carcinogenesis and the repression of transposable elements. (A transposable element is a DNA sequence that can change its position within a genome, altering the cell's genetic identity.)

MicroRNA (miRNA) is a small non-coding RNA (ribonucleic acid) molecule in plants, mammals and some viruses. Its functions include helping cells fight off RNA viruses and regulating the gene's production of proteins or RNA.

Professor Taguchi employs his cutting-edge methods to analyse complex situations involving heart failure, post-traumatic stress disorder (PTSD), acute lymphoblastic leukaemia (ALL), diabetes, renal carcinoma, and cirrhosis. In each case, there is a copious number of features of interest, but only a very small number of available samples.

The author demonstrates that PCA-based unsupervised feature extraction is a powerful method, when compared to other machine learning techniques. He also points out that in order to fully benefit from the technique, the user requires a deep understanding of the data to be analysed.

Professor Taguchi's book is a valuable resource for students, researchers, academics and practitioners.

TD-based unsupervised feature extraction is a more recent development. Professor Taguchi has applied it to many of the problems that he has already analysed using PCA-based unsupervised feature extraction. He also mentions its new application to *in silico* drug discovery.

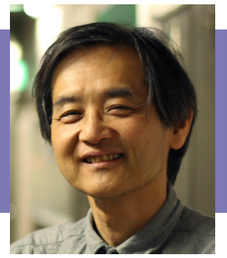
ADVANTAGES OF UNSUPERVISED FEATURE EXTRACTION METHODS

Professor Taguchi explains that when compared with conventional supervised methods, his unsupervised methods are more robust. Both PCA-based unsupervised feature extraction and TD-based unsupervised feature extraction avoid the overfitting that can easily occur when supervised methods stick too closely to the training data and cannot generalise enough to reliably deal with unseen data. In addition, he points out that unsupervised methods are independent of labelling and therefore more stable. Furthermore, in both his PCA and TC methods, the interaction between variables is considered before, rather than after the features are selected.

BROADER IMPACTS

Professor Taguchi's methods can be applied to a wide range of large p small n problems enabling his readers to analyse data sets with small samples and many features. The data features are easily interpreted as the output can be traced back to the various input features. The methodology, based upon linear algebra, offers a fast algorithm for analysing big data with output that is easily interpreted.

PCA and TD-based unsupervised feature extraction methods are powerful tools in the study of biological problems involving biomarker identification, gene expression, and drug discovery as demonstrated in this book and Professor Taguchi's numerous high-quality journal publications. This book is a valuable resource for students, researchers, academics and practitioners. Even though the applications demonstrated in the book focus on bioinformatics, the methodological framework is universal and can be employed in other areas of data science.



Behind the Research

Professor Y-h Taguchi

E: tag@granular.com **T:** +81-3-3817-1791 **W:** https://researchmap.jp/Yh_Taguchi/?lang=en

Research Objectives

Professor Taguchi proposes a new system of unsupervised learning for machine learning which would allow the answers for massive data sets to be predicted.

Detail

Y-h Taguchi
Department of Physics
Chuo University
1-13-27 Kasuga
Bunkyo-ku
Tokyo 112-8551
Japan

Bio

Professor Taguchi has been a physics professor at Chuo University, Tokyo, Japan since 2006. He received his Dr. Sci. from Tokyo Institute of Technology. He was an assistant professor there and moved to Chuo University in 1997. He was also a visiting professor at the European Bioinformatics Institute from 2009 to 2010.

Funding

Japan Society for the Promotion of Science

References

Taguchi, Y.H. (2020). *Unsupervised Feature Extraction Applied to Bioinformatics A PCA Based and TD Based Approach*, Springer, Switzerland. doi:10.1007/978-3-030-22456-1. Available at: https://doi.org/10.1007/978-3-030-22456-1 [Accessed 6th April 2020]

Personal Response

What initially inspired your research into unsupervised feature extraction?

"We human beings can have only limited intelligence, since we are not God. Thus, any kind of supervised learning is limited by our own intelligence. Nevertheless, a data-driven approach can overcome our intelligence, since we do not make use of pre-knowledge.

One day, when I was testing numerous supervised learning methods, I used PCA accidentally. I was surprised that it could tell me what I could never have predicted. I realised that if we can make further use of data-driven approaches, we can go beyond our own intelligence, since data is something like God. In this sense, an unsupervised approach is the closest method to God of data."

Yurchanka Siarhei/Shutterstock.com