SpringerBriefs in Computer Science

For further volumes:
http://www.springer.com/series/10028

David B. Skillicorn

Understanding High-Dimensional Spaces

Springer

David B. Skillicorn
School of Computing
Queen's University
Kingston, ON
Canada

ISSN 2191-5768 ISSN 2191-5776 (electronic)
ISBN 978-3-642-33397-2 ISBN 978-3-642-33398-9 (eBook)
DOI 10.1007/978-3-642-33398-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012948185

ACM Computing Classification (1998): H.2, E.1, I.2

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

High-dimensional spaces arise naturally as a way of modelling datasets with many attributes. Such a dataset can be directly represented in a space spanned by the attributes, with each record of the dataset represented as a point in the space with its position depending on its attribute values. Such spaces are not easy to work with because of their high dimensionality: our intuition about space is not reliable, and measures such as distance do not provide as clear information as we might expect.

High-dimensional spaces have not received as much attention as their applications deserve, partly for these reasons. Some areas where there has been substantial research are: images and video, with high-dimensional representations based on one attribute per pixel; and spaces with highly non-convex clusters. For images and video, the high dimensionality is an artifact of a direct representation, but the inherent dimensionality is usually much lower, and easily discoverable. Spaces with a few highly non-convex clusters do occur, but are not typical of the kind of datasets that arise in practice.

There are at least three main areas where complex high dimensionality and large datasets arise naturally. The first is data collected by online retailers (e.g. Amazon), preference sites (e.g. Pandora), social media sites (e.g. Facebook), and the customer relationship data of all large businesses. In these applications, the amount of data available about any individual is large but also sparse. For example, a site like Pandora has preference information for every song that a user has listened to, but this is still a tiny fraction of all of the songs that the site cares about. A site like Amazon has information about which items any customer has bought, but this is a small fraction of what is available.

The second is data derived from text (and speech). The word usage in a set of documents produces data about the frequency with which each word is used. As in the first case, all of the words used in a given document are visible, but there are always many words that are not used at all in it. So such datasets are large (because easy to construct), wide (because languages contain many words), and sparse (because any document uses a small fraction of the possible words).

The third is data collected for a security, defence, law enforcement or intelligence purpose; or collected about computer networks for cybersecurity. Such

datasets are large and wide because of the need to enable as good solutions as possible by throwing the data collection net wide. This third domain differs from the previous two because of greater emphasis on the anomalous or outlying parts of the data rather than the more central and common place.

High-dimensional datasets are usually analyzed in two ways: by finding the set of clusters they contain; or by looking for the outliers—almost two sides of the same coin. However, these simple strategies conceal subtleties that are often ignored. A cluster cannot really be understood without seeing its relationships to other clusters "around" it; and outliers cannot be understood without understanding both the clusters that they are nearest to, and what other outliers are "around" them. The development of the idea of local outliers has helped with this latter issue, but is still weak because a local outlier is defined only with respect to its nearest non-outlying cluster.

In this book we introduce two ideas that are not completely new, but which have not received as much attention as they should have, and for which the research results are partial and scattered. In essence, we suggest a new way of thinking about how to understand high-dimensional spaces using two models: the *skeleton* which relates the clusters to one another, and *boundaries in empty space* which provides a new perspective on outliers, and on outlying regions.

This book should be useful to those who are analyzing high-dimensional spaces using existing tools, and who feel that they are not getting as much out of the data as they could; also their managers who are trying to understand the path forward in terms of what is possible, and how they might get there. The book assumes either that the reader has a reasonable grasp of mainstream data mining tools and techniques, or does not need to get into the weeds of the technology but needs a sense of the landscape. The book may also be useful for graduate students and other researchers who are looking for open problems, or new ways to think about and apply older techniques.

Acknowledgments

My greatest debt is to Mike Bourassa. Discussions in the course of his doctoral research first surfaced many of the ideas described here, and we had many long conversations about what it meant to be interesting, before we converged on the meaning that is expanded here. I am also grateful to all my students who, by providing an audience for me to explain both simple and complex ideas, help me to understand them better.

Kingston, April 2012 David Skillicorn

Contents