ISSN 2186-7437

NII Shonan Meeting Report

No. 2013-4

Dimensionality and Scalability

Michael E. Houle
Vincent Oria
Arthur Zimek

May 20–23, 2013

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Dimensionality and Scalability

Organizers:
Michael E. Houle (National Institute of Informatics, Japan)
Vincent Oria (New Jersey Institute of Technology, USA)
Arthur Zimek (Ludwig-Maximilians-Universität München, Germany)

May 20–23, 2013

Description of the Meeting

Background

For many fundamental operations in the areas of search and retrieval, data mining, machine learning, multimedia, recommendation systems, and bioinformatics, the efficiency and effectiveness of implementations depends crucially on the interplay between measures of data similarity and the features by which data objects are represented.

When the number of features (the data *dimensionality*) is high, similarity values tend to concentrate strongly about their means, a phenomenon commonly referred to as the *curse of dimensionality*. As the dimensionality increases, the discriminative ability of similarity measures diminishes to the point where methods that depend on them lose their effectiveness.

One fundamental task, arising in applications of multimedia, data mining and machine learning, and other disciplines, is that of content-based similarity search. For such applications, features are often sought so as to provide the best possible coverage across a range of anticipated queries. However, for any given query, only a relatively small number of features may turn out to be relevant. When the dimensionality is high, the errors introduced into similarity measurements by the many irrelevant feature attributes can completely overwhelm the contributions of the relevant features.

The way in which the curse of dimensionality manifests itself varies from discipline to discipline. For example:

- In multimedia applications, direct representations of images and video frames typically involve hundreds or thousands of features. The visual vocabularies currently being sought for large image corpuses or video archives can have millions of visual words. These huge feature set sizes can lead to severe problems for clustering, classification, and any other applications that require content-based similarity search.

- Data mining and analysis applications often require the identification of relatively small clusters of mutually-similary data objects (*nuggets*). Identification of nuggets usually entails the generation of lists of objects similar

to query objects, to serve as candidates for cluster membership. Traditional clustering methods may be inadequate in identifying nuggets; even subspace clustering methods that specifically target groupings based on subsets of the feature set may be defeated by the sheer number of possible feature combinations.

To support operations in such areas, feature selection and other dimensional reduction techniques from machine learning are often considered in an attempt to improve the discriminability of similarity measures, and the scalability of methods that depend upon them. Yet even here, the complexity of searching through combinations of features can be prohibitive.

Dimensionality and Data Modeling

Many researchers and practitioners from different areas who are specifically working on problems involving the selection of features tend to be aware of the difficulties involved with high-dimensional data settings. Researchers in other areas are generally aware that the performance of their solutions depends on the dimensionality of their data sets, but are often not clear as to why.

In general, researchers tend to evaluate the inherent difficulty of tasks in terms of characteristics of the data set, such as the number of records, the number of feature dimensions, or the size of the data in bytes. However, the usual characterizations are not always indicative of the true difficulty involved in the processing of tasks for that set. For any given task performed on two different sets with the same numbers of records, features, and bytes, the performances could vary tremendously. On the other hand, data sets that are difficult to process for tasks in one domain tend to be difficult to process for other tasks in other domains.

Over the past decade or so, new characterizations of data sets have been proposed so as to assess the performance of particular methods. Such characterizations include estimations of distribution, estimation of local subspace dimension, and measures of intrinsic dimensionality of data. Although the applications affected by the curse of dimensionality vary widely across research disciplines, the characterizations and models of data that can be applied to analyze the performance of solutions are very general. Researchers from different disciplines, even though they work with different types of data for very different purposes, can nevertheless make use of a common set of data models and data characterizations. With a common framework in place for the assessment of the complexity of data sets, general techniques developed within one discipline for common subtasks such as search, clustering, classification, matching, and feature selection could conceivably be assessed for their applicability to tasks from other disciplines. Unfortunately, across the various disciplines, most such characterizations are as yet either unknown or underused, and techniques developed within one discipline for coping with the curse of dimensionality are constantly being reinvented by researchers in other disciplines.

Objectives

The goal of this meeting was to bring together researchers and students active in the areas of databases, data mining, pattern recognition, machine learning,

statistics, multimedia, bioinformatics, visualization, and algorithmics who are currently searching for effective and scalable solutions to problems affected by the curse of dimensionality. In particular, the objectives were:

- To survey the existing approaches used in dealing with the curse of dimensionality in these various disciplines, identifying their commonalities, strengths and limitations.
- To identify promising general data characterizations useful for the design and analysis of important subtasks common to these disciplines that are strongly affected by the curse of dimensionality, including (but not limited to) search, classification, clustering, and feature selection.
- To promote the adoption of theoretical analysis by researchers within these various disciplines.

Participants were not expected to give presentations of fully completed research. Instead, emphasis was placed on group input into the development of data characterizations, and the identification of future directions for research on dimensionality and scalability.

Participants

Laurent Amsaleg, IRISA-CNRS, France

Ira Assent, Aarhus University, Denmark

Ricardo J. G. B. Campello, University of Sao Paulo, Brazil

Peter D. Eades, University of Sydney, Australia

Vladimir Estivill-Castro, Griffith University, Australia

Björn Þór Jónsson, Reykjavik University, Iceland

Peer Kröger, Ludwig-Maximilians-Universität München, Germany

Michael Nett, NII, Tokyo, Japan

Chong-Wah Ngo, City University of Hong Kong, China

Miloš Radovanović, University of Novi Sad, Serbia

Jörg Sander, University of Alberta, Edmonton, Canada

Shin'ichi Satoh, NII, Tokyo, Japan

Koji Tsuda, AIST, Tokyo, Japan

Takeaki Uno, NII, Tokyo, Japan

Takashi Washio, ISIR, Osaka University, Japan

Pavel Zezula, Masaryk University, Brno, Czech Republic

Overview of Talks

The first part of the meeting comprised survey talks to illuminate the topic of "Dimensionality and Scalability" from different perspectives.

Scalability of Similarity Searching

Pavel Zezula, Masaryk University, Brno, Czech Republic

Similarity searching has been a research issue for many years, and searching has probably become the most important web application today. As the complexity of data objects grows, it is more and more difficult to reason about digital objects otherwise than through the similarity. In the presentation, we first discuss on concepts of similarity and searching in light of the contemporary Big Data problem before introducing a short survey of similarity search history. We analyze the bottlenecks of application development and discuss perspectives of search computing for future applications. Specifically, the directions towards similarity search cloud services and self-organizing search systems are outlined.

Obstacles and Some Remedies in High Dimensional Particle Filtering

Takashi Washio, ISIR, Osaka University, Japan

This review presentation focused on the impact on the curse of dimensionality to Monte Carlo based Bayesian estimation and learning, particularly, Particle Filtering (PF). This problem contains common issues over various Monte Carlo and Bayesian estimation problems.

First, concentration of measure on high-dimensional structures was discussed as one of the major effect of the curse of dimensionality. Some intuitive explanation of this effect and its mathematical analysis were provided.

Second, basic principles and some techniques of the particle filtering were outlined, and an important technical issue named degeneration of the particle weights is introduced. Furthermore, the impact of the concentration of measure on high-dimensional structures to the Bayesian estimation in the particle filtering was explained together its mathematical analysis. The review showed two main obstacles to the high dimensional particle filtering; (1) the statistical instability by the severe degeneration of the particle weights and (2) the large estimation bias by the concentration of measure.

Third, the review explained three remedies to the obstacles; (A) Rao-Blackwellised particle filtering techniques, (B) use of Markov Chain Monte Carlo (MCMC) for generating a posteriori distribution and (C) use of optimization with proposal distribution. (A) relaxes the obstacles (1) and (2) by reducing variables in PF computation, however its applicability is very limited to some special cases. (B) relaxes the obstacle (1) by not using particle weighting but Metropolis Hastings algorithm, however much computation is needed to obtain its sufficient effect. (C) relaxes the obstacles (2) by searching the maximum a posteriori probability, but no systematic framework has not been establish yet.

Finally, the review pointed out the needs of much more studies in this field to address many aforementioned issues.

Subspace clustering

Ira Assent, Aarhus University, Denmark

Clustering is an established data mining technique for automatically grouping objects based on mutual similarity. As we face increasingly high-dimensional data, effects attributed to the "curse of dimensionality" mean that in high-dimensional spaces, traditional clustering methods fail to identify meaningful clusters.

In this talk, I provide an overview over some of the approaches that have been proposed in the data mining community. Dimensionality reduction techniques aim at identifying a projection of the data in which most information is retained, while noisy attributes are removed. As this necessarily requires a global view on the relevance of attributes for all objects in the data set, subspace and projected clustering methods have been suggested as an alternative. In a nutshell, the idea is to identify the clusters along with their (locally) relevant subset of dimensions.

Different approaches to identifying clusters in subspaces have been proposed. Subspace search is a two-phase process, where the first phase identifies promising subspaces based on some interestingness criterion, and the second phase then subjects these interesting subspaces to traditional clustering algorithms. Subspace or projected clustering methods aim at identifying subspaces and clusters at the same time. Bottom-up algorithms work in an apriori-style manner, using low dimensional results to build candidates on higher dimensional dimensions in the subspace lattice. This may lead to impractical runtimes for large and very high-dimensional data sets, so grid-based methods partition the space either using a fixed layout or data-based hyperplanes. Alternatively, some subspace clustering algorithms avoid bottom-up search and instead search in a best-first manner, which may be supported by specialized data structures. To scale even more, approximate algorithms combine interesting regions in low dimensional spaces to region candidates where the number of attributes is much higher, allowing to skip many relatively low dimensional subspaces in-between.

Finally, I will give some examples of application areas such as time series or document mining, where high-dimensionality also is an issue, but where special properties of the data may allow more specific solutions to the problem.

Outlier Detection in High-Dimensional Data

Arthur Zimek, Ludwig-Maximilians-Universität München, Germany

High dimensional data in Euclidean space pose special challenges to data mining algorithms. These challenges are often indiscriminately subsumed under the term "curse of dimensionality", more concrete aspects being the so-called "distance concentration effect", the presence of irrelevant attributes concealing relevant information, or simply efficiency issues.

Albeit the infamous "curse of dimensionality" has been credited for many problems and has indiscriminately been used as a motivation for many new approaches, we should try to understand the problems occurring in high dimensional data in more detail. For example, there is a widespread mistaken belief that every point in high dimensional space is an outlier. This – misleading, to say the least – statement has been suggested as a motivation for the first

approach specialized to outlier detection in subspaces of high dimensional data, recurring superficially to a fundamental paper on the "curse of dimensionality". We will show in this talk that it is not as simple as that. We go into detail of some effects of the "curse of dimensionality" and discern truths and widespread misconceptions regarding high dimensional data with respect to outlier detection. Important aspects are

1. the concentration of distances
2. the effect of different proportions of irrelevant attributes
3. numerical discrimination of outlier scores vs. ranking of outlier scores
4. combinatorial issues (data snooping)
5. hubness

We discuss these effects and their consequences for outlier detection.

Scalability Problems in Bioinformatics

Koji Tsuda, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

In the last 10-15 years there has been a great increase of interest in space-efficient (succinct) data structures that are compressed up to the information theoretic lower bound. Compared to pointer-based naive data structures, the memory usage can be smaller up to 20-30 fold. I briefly present basics of succinct data structures and our recent work of indexing 25 million chemical graphs for similarity search in memory. In addition, fast all pairs similarity search methods called SketchSort and SlideSort are discussed.

Information Visualization

Peter D. Eades, University of Sydney, Australia

This is an introduction to Information Visualization. The talk begins with the purposes of visualization. Then a general model for the visualization process is described, as well as some example methods. We describe methods for evaluation of visualizations, and finish with a list of challenges for the future.

Scalability in Multimedia Search and Mining

Chong-Wah Ngo, City University of Hong Kong, China

In large-scale multimedia search, there are a large number of features in high dimensional space to be dealt with. Even an image can contain hundreds of local feature points. For a million-scale multimedia dataset, the visual search can run forever if without a proper indexing scheme to organize the features. This talk presents the indexing structures that have been successfully applied and adapted for this problem, and discusses issues such as feature dimensionality, memory consumption and query processing time. While searching against a million-scale

dataset is feasible, mining visual objects is an extremely difficult task due to the scalability problem. This talk presents some techniques addressing this problem, discussing reasons that there is yet to have an effective and efficient algorithm for multimedia mining.

Dimensionality and Scalability Issues in Multimedia

Shin'ichi Satoh, NII, Tokyo, Japan

Dimensionality and scalability issues in multimedia are discussed. Special emphasis was laid on image among other multimedia data, and also image search and image semantic classification were investigated as typical tasks. Global feature and local feature, aggregation strategies of local feature such as bag of visual word were discussed. Most cases features are high dimensional and sparse. Especially in image classification task, the number of possible classes introduces another direction of scalability issues.

How to Refine Pattern Mining

Takeaki Uno, NII, Tokyo, Japan

Pattern mining is one of the fundamentals in data mining. The aim of the pattern mining is, usually, to find all patterns of a specified class, in the given database. The problem has a long history of up to 20 years, and has been intensively studied. There have been proposed many problems and algorithms. However, pattern mining algorithms are no so much used in real-world applications. One of the reasons is that the number of solutions is usually huge, and many similar patterns appearing in similar places are found. Constraints given by background knowledge are often used to reduce the number, but this does not have any mechanism of automatically avoiding huge number of solutions. In this talk, we introduce a new concept of data properlization. Data properlization adds the lost edges and deletes unnecessary edges from the data so that the data will have much more clarity, that is, objective structures will have no ambiguity. For example, data properlization changes pseudo clique corresponding to a cluster to a clique. By applying the data properlization, the number of cliques will be drastically reduced, and thus we can efficiently find all local structures without implicit duplications. This approach would be the first result in data mining that has a mechanism of reducing the number of solutions, with simple and efficient computation.

Intrinsic Dimensionality and Discriminability of Data

Michael E. Houle, NII, Tokyo, Japan

For many large-scale applications in data mining, machine learning, and multimedia, fundamental operations such as similarity search, retrieval, classification, clustering, and anomaly detection generally suffer from an effect known as the 'curse of dimensionality'. As the dimensionality of the data increases, distance values tend to become less discriminative, due to their increasing relative concentration about the mean of their distribution. For this reason, researchers

have considered the analysis of structures and methods in terms of measures of the intrinsic dimensionality of the datasets. This presentation introduces a generalization of a discrete measure of intrinsic dimensionality, the expansion dimension, to the case of continuous distance distributions. This notion of intrinsic dimensionality of a distribution can be shown to precisely coincide with a natural notion of the indiscriminability of distances and features. Furthermore, for any distance distribution with differentiable cumulative density function, a fundamental relationship exists between probability density, the cumulative density (cumulative probability divided by distance), intrinsic dimensionality, and discriminability.

Summary of Discussions

During the second part of the workshop, the participants discussed some topics in different groups.

Intrinsic Dimensionality w.r.t. (Subspace-)Clustering and Outlier Detection

A group including Ira, Arthur, Jörg, Björn Þór, Ricardo, Peer, Miloš, Michael Houle, and Laurent focused on thinking about what the continuous form of the Intrinsic Dimensionality (ID) could bring for helping to address sub-space clustering and/or outlier detection.

The general understanding is that ID can be estimated/measured from a local point of view, that is by analyzing the vicinity of a specific point in space. Then, the knowledge acquired from many points can be aggregated one way or the other in order to return a more global information of the ID of the entire dataset.

The ID around a point can be measured, increasing the radius used for the analysis. This, in turn, gives a curve. A similar operation can be done to plot a Φ-curve instead of an ID-curve. One direction to follow in order to get a better understanding of the entire process is to plot several such curves using real datasets (such as the ALOI dataset) as well as synthetic ones. Note that determining good ways for generating such datasets is not fully clear; yet, simple datasets such as the ones mentioned by Arthur during his talk as well as the typical datasets addressing classical clustering problems might be used. This well controlled environment might help discovering what is at stake with Φ- and ID-curves.

Several interesting point configurations were discussed, as they might highlight specific properties of the ID stories. Three such configurations have been discussed: assuming there is a cluster in space, how curves from two close points within the cluster relate, from two points in the same cluster but far apart, and one point inside the cluster and another outside. As curves rely on a metric, it might be worth studying the impact of using regular distances, ranks, the number of shared neighbors. Plotting curves might give insights, as well as aggregating them into some kind of surface, possibly showing new properties (local min, max, ...).

While computing Φ- and ID-curves encompasses all the dimensions, it is possible that few features only might define a subspace that is hard to capture

from the global point of view of the ID definition. No solution to this is foreseen, except to try to build on this in order to validate which features to pick to reach the observed ID — but this idea remains controversial and unclear. Furthermore this might only identify axis-parallel subspaces. A PCA based mechanism might be needed to cope with non aligned subspaces.

Another direction is to investigate what happens when we use these new notions in existing works such as DBSCAN, or density-based clustering and outlier detection in general. Of course, that study should include analyzing what happens in low-dimensional (possibly little or no impact for ID) vs. high-dimensional data (possibly higher impact).

Note that some current work using a concentration measure might have strong relationships with estimating the ID.

Note also that checking the behavior of the ID when getting close to the borders of the space is probably required. This is indeed a concern as this might alter the number we observe w.r.t. the numbers we expect to observe, as a part of space by definition can not contain any point while radii are oblivious to this.

Computation & Estimation of Intrinsic Dimensonality

A group including Takeaki Uno, Koji Tsuda, Takashi Washio, Peter Eades, and Michael Nett discussed some implications from a machine learning persepctive.

Support vector machines using radial base function kernels perform well for many practical tasks. The underlying kernel distance is parameterized in the kernel width parameter. Usually, this parameter is chosen by supervised preprocessing methods like cross-fold validation. However, a property that determines particularly well-performing parameter choices is yet to be discovered. Nevertheless, we suspect that such a property may be related to the notion of discriminability (and equivalently intrinsic dimensionality). Unfortunately, as it is not well understood in the community what properties of the kernel distance distribution are important, we are likely forced to discover or learn such properties empirically through extensive experimentation.

As a second topic, we discussed how the decomposability of the distance distribution within the model of intrinsic dimensionality proposed by Michael Houle may be marginalized. While it is not at all clear how to interpret the marginalized contribution of probability measure of individual points in space, it indicates that this issue is linked to the question of how intrinsic dimensionality and distance distributions between different points are related.

Intrinsic Dimensionality and Graphs

A group including Vlad, Koji, Takashi, and Peter formulated research questions related to graphs:

Such questions are:

- What is the relationship between ID and the "expansion" of a graph?
- What is the ID of a social network (or a scale-free network)
- How does ID vary with changing distance metrics?

An idea was to produce some stories in low dimensions with small data to illustrate ID.

A conjecture: "A graph with low ID has a good drawing."

Intrinsic Dimensionality and Empirical Results

A group including Peter, Vlad, Koji, and Björn Þór discussed questions regarding the relationship between theory and empirical results.

We need an underlying theory to explain our empiricial results. Such a theory might be developed around the concept of intrinsic dimensionality. For example, something should explain the trade-off between increasing dimension and paying penalty in search computational resources

A second issue is the need for better benchmark data sets for research into the applications where large scale is required. Papers which describe methods aimed at large data but only tested on small data are not valuable. Ideas to encourage the development and curation of benchmark data sets could be to have a special issue of a journal on benchmark data sets. More generally, publication of methodology should be encouraged. Other possibilities might be competitions and particular workshops. Questions are if it might be possible to raise funding for data curation? To publish benchmark data might also involve copyright issues and privacy issues.

Dimensionality and Multimedia

The discussion group on dimensionality and multimedia included Chong-Wah Ngo, Vincent Oria, Shin'ichi Satoh and Pavel Zezula. Michael Nett joined the group on the last day.

The complexity of next-generation retrieval systems originates from the requirement to organise massive and ever growing volumes of heterogeneous data and meta-data, together with the need to co-ordinate scalable management of similarity matching. The problem starts with data acquisition of weakly structured or completely unstructured data, such as images and video, which necessarily need innovative techniques for information extraction and classification to increase their findability. Raw multimedia data cannot be directly used when performing tasks such as search and classification due to the amount of data involved, and the lack of interpretability. The content of multimedia data is often represented as high dimensional vectors called descriptors that depend on the media type and the application. For images, there are mainly two types of descriptors: global descriptors such as color histograms, and local descriptors such as SIFT. Local features are often aggregated to form visual words with each image represented by a bag of visual words (BoVW). With the BoVW approach, an image is represented by a sparse vector in a high-dimensional space (which can be in excess of one million dimensions). Dimensionality and scalability issues in image databases arise due to the number of dimensions used to represent the visual features, the number of local features per image, the number of visual words, the number of queries and the number of users.

In principle, we consider search and object findability as two principal and synergistic aspects of information access. They both represent the effectiveness and efficiency challenges which need innovative theories and technologies, and must be studied together to allow the development of qualitatively new retrieval

tools in the future. The objective is foundational in nature in that it addresses the theoretical limits of similarity retrieval in the context of the Big Data problem — a development of scalable solutions with low intrinsic dimensionality is essential. The research issues related to dimensionality and scalability that we have identified are:

- Local effective feature selection: the multimedia field is currently dealing with some of the most difficult datasets. Face data is a typical example where current similarity functions cannot discriminate among data points. One reason for the lack of discriminability is that some of the dimensions used for representing the features can be useless in the locality of the query. A research direction that we are proposing to investigate is to locally identify and select the most effective features to represent the data. The research tasks involved are:
  - Data representation (how to represent the database with only selected local effective features and what type of similarity function will support the data representation).
  - Adaptive search (adapt the search to each query point since the data representation will depends on the query point neighborhood).
  - Indexing (which index structure and algorithms will properly support adaptive search and local effective features).
  - Machine learning (How to learn only local effective features using a machine learning approach).
- Annotation and scalability: it has become feasible to learn large number of concept classifiers, with a large number of labeled images (e.g. ImageNet) and weakly labeled data freely available. The issue of scalability becomes a concern when considering (1) how to effectively learn a new classifier, and (2) how to efficiently annotate a sample against a large pool of classifiers. For (1), an interesting topic is the leveraging of the available classifiers for learning a new classifier that can scale in terms of learning speed and classification robustness. For (2), organizing the classifiers hierarchically can naturally speed up the annotation, but also result in performance drop compared to testing against all classifiers. Therefore, there is a tradeoff between annotation accuracy and efficiency. An interesting topic is to identify the trade-off, taking into account the organization of classifiers and the reliability of classifiers.
- Dimensionality and similarity function: how does dimensionality affect different similarity functions, and how to design similarity functions that can cope with high dimensionality?
- Training and dimensionality: is training really needed in the visual word creation? Training in high dimensional spaces may cause overfitting, and subsequently a cross-domain problem. In this case, randomly selected local features to be used as codebook can yield better results.