ISSN 2186-7437

NII Shonan Meeting Report

No. 2015-9

Dimensionality and Scalability II: Hands-On Intrinsic Dimensionality

Laurent Amsaleg
Michael E. Houle
Vincent Oria
Arthur Zimek

June 29–July 2, 2015

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

Dimensionality and Scalability II: Hands-On Intrinsic Dimensionality

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

June 29–July 2, 2015

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. The effects of the curse of dimensionality on searching/clustering methods are well known and well documented.

In turn, many methods have been invented to delay the effects of the curse of dimensionality. This includes dimensionality reduction and feature selection methods. These methods work to some extent, but the fundamental problem which is the indiscriminability among the data points remains.

Over the past decade or so, new characterizations of data sets have been proposed so as to assess the easiness of data sets. 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. In turn, quite similar data models and data characterizations have been invented by researchers from different disciplines. Unfortunately, researchers from one domain typically ignore what researchers from other, or even close domains have invented. It is indeed the case for the new tools helping characterizing high-dimensional data sets.

2013: NII Shonan Meeting #1–Dimensionality and Scalability

In May 2013, a NII Shonan meeting was held 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 looking for effective and scalable solutions to problems caused by the curse of dimensionality. The objectives of this workshop were essentially to survey current approaches proposed to deal with the curse of dimensionality in different disciplines, identifying their commonalities, strengths and limitations; to clarify the potential impact of such approaches on core tasks such as search, classification and clustering.

During four days, 16 participants brainstormed, identifying future directions for research on dimensionality and scalability. Ten survey talks helped the attendance to better understand the multiple facets of the problems. The active discussions quickly identified intrinsic dimensionality estimators as a urgent need. Further discussions in small working groups focused on various topics in order to clarify the interplay between intrinsic dimensionality and clustering and outlier detection, multimedia, graphs, feature selection, etc.

See the report on this meeting under:
`http://www.nii.ac.jp/shonan/wp-content/uploads/2011/09/No.2013-4.pdf`

2014: a One Day Seminar on Dimensionality and Scalability

Since that workshop, the people who participated to the first Shonan meeting continued working on these topics and several key results were obtained. The outcomes of these researches were presented at a special one day seminar organized in March 2014 at NII.The participants to the one day seminar was a small subset of the participants to the first Shonan Meeting.

On the theoretical side, using the theory of extreme values enabled the specification of various intrinsic dimensionality estimators and their implementation. In parallel, several contributions allowed to better understand the impact of the intrinsic dimensionality of data sets on the quality of similarity searches and the underlying indexing techniques. Furthermore, some methods allowing to detect outliers based on intrinsic dimensionality were proposed. Additionally, exploiting k-nearest-neighbors graphs as well as carefully normalizing scores and distances for ensemble methods were proposed as remedy to the dimensionality damages.

Overall, that day was very useful and allowed to share the newest results and to identify possible research directions and collaborations. Briefly outlined, four major perspectives delineate the foreseen research agenda: deeper investigation of the complicated relationships between intrinsic dimensionality and high-dimensional indexing/clustering, clarification of the connection between shared nearest neighbors, hubness and intrinsic dimensionality, tackling outlier detection with intrinsic dimensionality estimators and finally better understanding the coupling of the notion of similarity to distance ensembles.

2015: NII Shonan Meeting #2–Hands-On Intrinsic Dimensionality

It was about time to organize a second edition of the NII Shonan workshop in order to (i) share the results obtained so far with the full crowd, (ii) leverage on the initial collaborations to consolidate research agendas and (iii) get feedback from new researchers concerned with the general problems of dimensionality and scalability.

The first goal was to share our latest discoveries by participants to the first Shonan meeting and make findings available the all the people of interest. We have now a much better understanding on intrinsic dimensionality and have had success in designing intrinsic dimensionality aware algorithms for search and clustering. This is a clear sign of the success of the first edition: much happened afterwards, in several research labs worldwide.

We now have estimators that can compute local values of dimensionalities, we also have operational computing rules making use of these estimators to prune search spaces, to filter noisy points and to enhance indexing and clustering. It is crucial to share this knowledge in order to make progress and to find other use-cases where such estimators can help defeating the curse of dimensionality problems.

Our second goal is to foster collaborations. The first Shonan workshop fully succeeded in bootstrapping collaborations that resulted into the contributions presented during the one day seminar. One goal of the second edition of the Shonan workshop is to push this further and not only consolidate existing collaborations, but to initiate new ones spanning the other relevant domains (machine learning, multimedia, bioinformatics, ...) in part represented during the first edition.

The third goal is to recruit more people. Compared to the first edition, we have gained in maturity on the theoretical side as well as on the practical side. It will be beneficial to disseminate these elaborated ideas to researchers concerned with dimensionality and scalability problems, beyond the participants to the first workshop.

Participants

Laurent Amsaleg, IRISA-CNRS, France

James Bailey, University of Melbourne, Australia

Oussama Chelly, National Institute of Informatics (NII), Japan

Michel Crucianu, Conservatoire National des Arts et Metiers (CNAM), France

Vladimir Estivill-Castro, Griffith University, Australia

Michael E. Houle, National Institute of Informatics (NII), Japan

Alfred Inselberg, Tel Aviv University, Israel

Ata Kaban, University of Birmingham, UK

Ken-ichi Kawarabayashi, National Institute of Informatics (NII), Japan

Peer Kröger, Ludwig-Maximilians-Universität München, Germany

Pei-Ling Lai, Southern Taiwan University of Science and Technology, Taiwan

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

Vincent Oria, New Jersey Institute of Technology (NJIT), NJ, USA

Srinivasan Parthasarathy, Ohio State University, USA

Miloš Radovanović, University of Novi Sad, Serbia

Shin'ichi Satoh, NII, Tokyo, Japan

Ansgar Scherp, Kiel University, Germany

Erich Schubert, Ludwig-Maximilians-Universität München, Germany

Mahito Sugiyama, ISIR, Osaka University, Japan

Kai-Ming Ting, Federation University, Australia

Nenad Tomašev, Google, USA

Takeaki Uno, NII, Tokyo, Japan

Takashi Washio, ISIR, Osaka University, Japan

Kaoru Yoshida, Sony Computer Science Laboratories, Japan

Pavel Zezula, Masaryk University, Brno, Czech Republic

Arthur Zimek, Ludwig-Maximilians-Universität München, Germany

Overview of Talks

The first part of the meeting comprised survey talks to illuminate the topic of "Intrinsic Dimensionality" (ID) from different perspectives.

An Extreme-Value-Theoretic Foundation for Similarity Applications

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 similarity applications in terms of measures of the intrinsic dimensionality (ID) of the data sets. This presentation is concerned with a generalization of a discrete measure of ID, the expansion dimension, to the case of continuous distance distributions. This notion of the ID of a distance distribution is shown to precisely coincide with a natural notion of the indiscriminability of distances, thereby establishing a theoretically-founded relationship among probability density, the cumulative density (cumulative probability divided by distance), intrinsic dimensionality, and discriminability. The proposed indiscriminability function is shown to completely determine an extreme-value-theoretic representation of the distance distribution. From this representation, a characterization in terms of continuous ID is derived for the notions of outlierness and inlierness of data.

Outlier Detection in High Dimensional Data based on Intrinsic Dimensionality

Arthur Zimek, LMU Munich, Germany

In this talk, we introduce a new method for evaluating local outliers, by utilizing the continuous intrinsic dimension (ID), which has been shown to be equivalent to a measure of the discriminative power of similarity functions. The proposed local outlier score, IDOS, uses ID as a substitute for the density estimation used in classical outlier detection methods such as LOF. An experimental analysis is provided showing that the precision of IDOS substantially improves over that of state-of-the-art outlier detection scoring methods, especially when the data sets are large and high-dimensional.

Unsupervised Feature Selection Using Local Intrinsic Dimensionality

Oussama Chelly, NII, Tokyo, Japan

As the dimensionality of data increases, the efficiency and effectiveness of various learning algorithms tends to degrade. In this paper, we propose new

filter approaches for unsupervised feature selection whose selection criteria assess the ability of features to discriminate within the neighborhoods of data points, according to a recent model of the local intrinsic dimensionality of continuous distance distributions. By ranking and selecting those features which are most discriminative under the model, our method seeks to improve the overall local discriminability of the distance measure within the data set. Experiments on several real-world datasets are conducted to compare the performance of our approach with state-of-the-art methods.

Intrinsic Dimensionality & Indexing Observed at Large Scale

Laurent Amsaleg, IRISA-CNRS, Rennes, France

In this talk, I am presenting several experimental results that are obtained when using the maximum likelihood estimation method against a very large benchmark created to evaluate nearest-neighbor (NN) search strategies. This benchmark contains the 1000 NN of each of 10000 different query points, the NNs being determined from a collection of 1 billion SIFT descriptors. Two series of results will be discussed. The first series is discussing the distribution of the IDs and of the distances to the NN, observed in this benchmark. The second series correlates the observed ID with the quality performance of 3 state-of-the-art high-dimensional indexing methods.

This work is very much in progress. The hope is to better understand the relationships between ID measures and the effectiveness of retrieval, which extends to close problems such as classification and outlier detection.

Hubs in Nearest-Neighbor Graphs: Origins, Applications and Challenges

Miloš Radovanović, University of Novi Sad, Serbia

The tendency of k-nearest neighbor graphs constructed from tabular data using some distance measure to contain hubs, i.e. points with in-degree much higher than expected, has drawn a fair amount of attention in recent years due to the observed impact on techniques used in many application domains. This talk will be organized into three parts: (1) Origins, which will discuss the causes of the emergence of hubs (and their low in-degree counterparts, the anti-hubs), and their relationships with dimensionality, neighborhood size, distance concentration, and the notion of centrality; (2) Applications, where we will present some notable effects of (anti-)hubs on techniques for machine learning, data mining and information retrieval, identify two different approaches to handling hubs adopted by researchers – through fighting or embracing their existence – and review techniques and applications belonging to the two groups; and (3) Challenges, which will discuss work in progress, open problems, and areas with significant opportunities for hub-related research.

Estimating First and Second Order Intrinsic Dimensionality

Oussama Chelly, NII, Tokyo, Japan

Estimating intrinsic dimensionality (ID) has several applications in machine learning, databases, and data mining. Developing better estimators of ID can improve search, classification, outlier detection, projection and feature selection. While global estimation methods (including topological, fractal and graph-based approaches) measure the dimensionality of an entire dataset, local methods measure the dimensionality around a specific point. Several estimators of local ID are proposed and analyzed based on extreme value theory, using maximum likelihood estimation (MLE), the method of moments (MoM), probability weighted moments (PWM), and regularly varying functions (RV). Global and local approaches were compared using both real and artificial data. Second order ID is a new concept that can be viewed as the normalized rate of change of local ID. It may be used as a criterion for 'inlierness'. Maximum likelihood estimation (MLE) and the method of moments (MoM) lead to non-closed-form estimators of second order ID. Experimental work is being conducted to validate these estimators.

Functionality for Intrinsic Dimensionality Available in the ELKI Framework

Erich Schubert, LMU Munich, Germany

In this short presentation, and overview of the implementations available for intrinsic dimensionality in the data mining framework ELKI was given. In order to make experiments with intrinsic dimensionality easier, this toolkit is available as open source to encourage use and contributions. The existing algorithms and visualization can be easily used to understand and study the estimators and their impact on algorithms, and the availability of nearest neighbor search indexes allows experimenting on larger data sets.

Summary of Discussions

Hubness

A group including Laurent, Miloš, Nenad, Arthur, Kai, Srinivasan, Erich, Kaoru, Pavel, Mahito, Takeaki, Takashi and Michael focused on discussing the issues related to the phenomenon of hubness and how it might be possible to improve the performance of various tasks and systems in data with high-hubness and high intrinsical dimensionality.

The consensus is that many standard approaches tend to be negatively affected by high data hubness, due to the asymmetry of implicit relevance of various data points, and the fact that the skewed distribution of inferred relevance is often misaligned with respect to the underlying relevance as perceived the user. This misalignment is due to the semantic gap between the high-level concepts and the low-level feature representations that are used in many applications. Therefore, novel and innovative approaches are required in order to better handle instance-based tasks in many-dimensional data. Several promising directions have been discussed, as well as several prominent use cases.

The potential influence of hubness on the stability and robustness of graph compression was determined to require closer examination. Graph compression is based on setting the appropriate thresholds in kNN-graphs or epsilon-graphs to perform sparsification and capture the underlying highly connected components/cliques. Past experiences suggest that shifting these thresholds results in phase transitions, making it difficult to select an appropriate threshold. As these transitions may be a result of the underlying local distance concentration and/or hubness, similarity-based hubness reduction was mentioned as a potential approach for dealing with this problem. In order to facilitate better understanding of the related issues, a dual examination of epsilon-graphs and kNN graphs with respect to the phase transitions in thresholding was proposed.

Hub duplication was mentioned as a strategy for dealing with hub-prone network clustering, and was also suggested as an approach for kNN graph mining. State-of-the-art sparsification methods for graphs were discussed that impact hubs more aggressively than other points. Local methods that take into account the neighborhood information were shown to perform better than global methods.

Mass-based density estimation was discussed in the context of extrinsic distance measures capable of better handling high-dimensional use cases. The distance between two points is expressed as the probability of the minimal partition covering the points in the current (rectangular) partitioning of the data. As such, it was shown to be superior to density when used for high-dimensional data. Examples were given for anomaly detection. It was suggested that comparisons to other known extrinsic measures that reduce hubness in the data should be done, including local scaling, non-iterative contextual dissimilarity, mutual proximity, shared-neighbor distances and hubness-aware shared neighbor distances.

Skewness of the neighbor occurrence frequency distribution was judged to be insufficient to capture all the relevant properties of the kNN graph related to hubness. The use of scale-free metrics was suggested as an alternative.

Distance ensembles and representation learning were mentioned in the context of working with multi-representational image data where different repre-

sentations or parts of a merged representation are formed by using different feature types that capture different notions of similarity in images. Image Hub Explorer was proposed as a platform for performing exploratory comparisons of the hubness-related properties of different image feature representations.

Scalable and robust kNN search is an important problem for Big Data. We have discussed the potential for implementing a scalable version of the hubness-aware shared-neighbor *simhub* distance measure in order to facilitate scalable hubness-aware metric learning on top of primary distances. Such an approach would have the potential to offer quality improvements on top of computational benefits. Hubness-aware re-ranking within the larger kNN context was judged to be promising, based on good hubness ratios. Transitive Hamming and Levenshtein distances were presented as a promising way to deal with various pattern recognition tasks. In transitive distances, element-wise mismatch is combined with the cost of transition to form the transitive dissimilarity scores. Dynamic programming implementations are available.

Feature Selection (VEC)

Most of this discussion drew inspiration from the earlier presentations by Oussama Chelly. The discussion emphasized the importance of feature selection in dealing with high-dimensional data, wherein unnecessary, redundant, or correlated attributes are removed. Such removal transforms the problem into one with fewer dimensions, reducing the resources required for data analytics, and improving the quality of the result. An important distinction is made with respect to feature extraction. Methods such as PCA may reduce the dimensionality but destroy the interpretation of the independent variables.

The main challenge in feature selection is in unsupervised settings. This is contrasted with supervised settings, in which class-labels assist in identifying those features that are meaningful for the classification. Many methods already exist for the supervised setting, starting from information gain approaches.

In the context of high-dimensional data sets, the methods that seem desirable are augmentative (forward) methods as opposed to backward methods; that is, it seems preferable to start with an initially empty feature set and to progressively select relevant features, as we expect earlier termination than in a backward approach starting with the set of all attributes, from which features are removed. The seminar discussions and presentations suggest that in computationally tractable data sets, the Local Intrinsic Dimension is usually orders of magnitude smaller then the representational dimension of the data, suggesting that if feature selection were to be optimal, only a very few features would likely be required.

This leads to the first challenge.

> *In forward methods for feature selection, the fundamental task is that of selecting a new feature to enlarge the current set of features. What are the best methods for achieving this?*

This was followed by the discussion on how Local Intrinsic Dimensionality (ID) was used earlier by Oussama to derive methods of feature selection.

The method was summarized as a ranking of all columns/features. Each column score is derived from the ID scores of all exemplars estimated as a 1-dimensional set. The Maximum Likelihood estimator for the ID is computed

(using 100 nearest neighbours). The score of a column is the highest value once the 5% highest values have been removed (declared as outliers).

This lead to 2 subquestions

1. Is cutting the top 5% local ID scores a good rule of thumb? Should this parameter be derived from the data set?

2. How should the values of each exemplar be aggregated to produce the score of the column?

In general, the elaboration of the general forward methods of feature selection offer another challenge:

> *How can one ensure that an added feature is not redundant or correlated with the earlier set?*

The challenge derives from the fact that testing for correlations seems expensive; however, see later summary of the discussion with visualization, the visualization methodology (and software) of Alfred Inselberg suggest that humans can rapidly detect correlations using parallel coordinate representations of data sets with large dimensions.

Oussama's approach modifies the scores of features that are candidates to be added to the selected set by increasing the weight of those features that have more points not already covered (where a point covers a feature if had low local-intrinsic dimensionality and the feature has already been chosen).

Because of the dependency on using a similarity measure in the approach above, another problem is the following.

> *How dependent (fragile/robust) are methods of feature selection to the similarity measure?*

And also,

> *How can local intrinsic dimensionality be used to select (or design) a similarity measure?*

Another puzzling aspect is that random selection of features seems particularly efficient, with little penalty in quality. To the practitioner, the forward method (already faster than the backward method) seems computationally expensive: many estimators of local intrinsic dissimilarity must be computed, and expensive tests must be performed to prevent the selection of redundant or correlated features.

> *How is then possible to demonstrate in a measurable way that feature selection in the forward approach is substantially better than a random choice?*

More importantly, random selection would be so fast that it could be reiterated many times, from which the best set of features could be selected as a final result.

There is an overarching theme that leads to a specific problem.

> *How can we develop a methodology for the evaluation of feature selection that firmly establishes that a suitable (or 'best') set of features has been selected?*

Derived from this,

> *How can one determine which specific elements of a feature selection process are responsible for the success of the result?*

Other discussions were more speculative, such as one on the topic of image databases, where each image is represented by a very large vector of features. Perhaps feature selection could be achieved by changing the domain to image description or tagging, and then using similarity measures with stronger semantics.

Finally, forward methods for feature selection are incremental hill-climbers. Further investigation can explore the potential of selecting features in batches. While this may imply a penalty in cost due to combinatorial explosion, there is still the following issue to investigate:

> *Relative to an optimum feature set of size S, what are the relative advantages and disadvantages of using incremental forward feature selection, as compared to pairs or triples of features?*

Proposals

There were two immediate proposals for action.

1. We should work/collaborate in generating interesting data sets that illustrate the challenges and benefits of feature selection methods.
2. We should generate a repository of interesting data sets and benchmarks for feature selection in the unsupervised case.

Similarity Measures

Discussion on this topic included James, Vincent, Ansgar, Michael, Michel, and Pavel.

Questions addressed in the discussion included:

- How can we deal with situations where multiple types of similarity exist?
- Is it preferable to have a feature set which captures just a single perspective? Is it reasonable to compute intrinsic dimensionality when the feature set contains a mixture of perspectives?
- Suppose we have multiple feature sets, each capturing a different perspective, how do we compute the ID? Can it be done in an ensemble or averaging manner?
- If we have multiple feature sets, could we warp an object so that it has same ID in each separate feature space? After warping, perhaps the similarities are more comparable for an ensemble?

Relevant object warping work has been performed by Michael Houle/Vincent Oria, targeting the scenario of label propagation in a graph. For an object we wish to determine i) a feature set and ii) its nearest neighbors. They have developed an alternating optimization style technique that compute this. This can be viewed as a type of warping for an object. What is the connection to ID?

Learning Guarantees

Discussion on this topic included Ata, Oussama, and Pei-Ling.

Machine learning is concerned with algorithms that are able to draw valid conclusions (i.e. generalise) from a finite number of training examples. The number of training points required for good generalisation is called sample complexity. Many machine learning methods scale badly with the data dimension, both in terms of sample complexity and computation time.

We started with a short presentation of some current work in progress by Ata, about nearest neighbour classification, in which the generalisation error and the associated sample complexity are expressed as a function of the intrinsic dimension (ID). This improves the known exponential (in dimension) sample complexity on the nearest neighbour learning rule to exponential in the ID. However, the notion of ID being used in this work is a global one.

Open questions:

1. How does the particular global ID used in this work (namely, the metric entropy integral) relate to the local ID as defined by Michael? Would it be possible to improve the current learning guarantee by using the local ID?

2. More generally, can ID (or perhaps an extension of the current form of this notion) help us characterise the number of training points the learning machines need to generalise? Intuitively, a learning problem should be easier if the ID of the data is lower than the observed dimension. Some generalisation bounds indeed depend on the observed dimension. Can we improve these bounds by using ID?

3. Directly related to the previous question, we need to pick apart various types of learning machines, e.g. supervised, unsupervised, etc, but also parametric, nonparametric, semi-parametric, etc. The ID of the input space, in the current 'unsupervised' sense of the notion of ID, will be relevant to some of these learning settings but not to others. Therefore, a further open problem is to define things like a 'supervised' ID - that is, ID relative to a specific task. This way we might be able eventually to relate ID with notions of large margin (and margin distribution) as currently used in the literature in statistical machine learning theory. (We can think of a linear classification problem as an $O(1/\text{margin}^2))$-dimensional problem.)

4. The previous open questions concerned sample complexity (training set as a resource). A complementary open problem is to identify / devise learning algorithms whose computational complexity scales with the ID of the data.

Visualization with Parallel Coordinates

The discussion was focused on the merits of parallel coordinates an an enabler of the human in the loop for analysis of large data set. It was proposed that the tremendous human capacity to spot patterns consists of a parallelization step where the data in the visualization is processed by the visual inspection 'in parallel'.

Debate about whether a 3D visualization of the parallel coordinates display of a high-dimensional data set is fruitful. Apparently the 2D view is to be preferred, as a paper by Johansen has already established. Nevertheless, a 2D display enables several patterns, in particular, correlations to be detected.

A demonstration as given by Alfred that illustrated even how inflection points in high dimension become transparent in parallel coordinates. Although that is a sophisticated example, linear dependencies become readily apparent to the eye trained with the methodology and the software for interaction.

The methodology includes a linear method to produce all permutations that display two independent variables next to each other in the visualization. Adjacent variables in the visualization enable the discovery of relationships between them upon inspection. It seems an open issue as to:

> *How to efficiently scroll though all possible triplets?*

Also interesting is to explore:

> *Why have methods such as decision lists and decision trees not used parallel coordinates for visualization?*

Proposals

1. To use the visualization methodology and interactive software in a forward algorithm of feature selection to at least visually examine the possibility of eliminating correlations or redundant variables when adding one feature incrementally in a forward feature-selection approach.

2. To provide visualizations of local-intrinsic dimensionality that could inform the process of feature selection.