ISSN 2186-7437

NII Shonan Meeting Report

No. 2018-13

Data Dependent Dissimilarity Measures

Kai Ming Ting
Takashi Washio
Ata Kaban

October 15–18, 2018

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

Data Dependent Dissimilarity Measures

Organizers:
Kai Ming Ting, Federation University Australia
Takashi Washio, Osaka University, Japan
Ata Kaban, Birmingham University, UK

October 15–18, 2018

This meeting aims to provide a forum to

- Discuss recent development in data dependent dissimilarity measures,
- Plan for future research directions in the next 2-5 years, and
- Establish research collaboration towards the research directions.

While the conventional data independent distance metric has been the primary means to measure dissimilarity of any two points in a given space, research in different fields has provided evidence that data dependent dissimilarity, where data distribution has the primary influence on the dissimilarity, is a better measure to find the closest match neighbourhood of a query—a core computation demanded in automated tasks such as classification, clustering, anomaly detection and information retrieval.

Advocates of data dependent dissimilarity include psychologists and computer scientists. Researchers in machine learning have advocated distance metric learning—a method which learns a mapping such that the mapped points are in the Euclidean space. In the supervised learning context, the mapping amounts to reducing the distance between points of the same class and increasing the distance between points of different classes in the mapped Euclidean space. It is also viewed as a way to learn a generalised (or parameterised) Mahalanobis distance, subject to some optimality constraint, from a dataset. Some data dependent dissimilarity measures, which require no learning, have been proposed, for instance, Mahalanobis distance, the term-weighted Cosine distance, cdf and rank transformations, and information theoretic definitions of similarity.

The need for data dependent dissimilarities came up in various different forms, implicitly or explicitly, in different subfields of machine learning and data mining. For instance, kernel methods, new definitions of similarity or dissimilarity for structured types of data, and the use of side information or 'privileged information' i.e., additional data available only at training time to inform the choice of metric to be used.

It is interesting to note that many existing data dependent dissimilarity measures are either metric or pseudo-metric. This is due to the following assumption: a necessary condition for the above mentioned automated tasks is that the dissimilarity measures must be a metric.

The psychological tests conducted in the 70s have shown that the dissimilarity between two instances, as judged by humans, is influenced by the context of measurements and other instances in proximity. It is suggested that a dissimilarity measure which is akin to one aspect of human's judged dissimilarity is: two instances in a dense region to be less similar than two instances of equal interpoint distance but located in a sparse region. In addition, the judged dissimilarity does not satisfy the metric constraints.

Recent research has provided more concrete evidence that nonmetric data dependent dissimilarity measures can be an effective alternative to distance metric to overcome the weaknesses of existing distance-based neighbourhood algorithms.

In the literature, the term 'data-dependent' has been used to mean different things:

(i) In the context of multiple kernel learning (e.g., [1]), the term means using a dataset to learn a weight for each user-defined data-independent kernel, in a (linear or non-linear) combination of multiple kernels in order to reduce the risk of choosing a bad kernel for the task at hand.

(ii) In the context of distance metric learning [2], the term means the use of class information and the training set to transform data f (usually accompanies a dimension reduction) to achieve the desired metric: $d(\mathbf{x}, \mathbf{y}) = ||f(\mathbf{x}) - f(\mathbf{y})||_2$.

(iii) In the context of conformal transformation [3], the term means modifying a data independent kernel to the class distribution of the data. Like distance metric learning, class information in the data plays a key role here. Similarly, RF kernel [4] produces a classifier from class-labelled data.

(iv) A kernel or similarity which depends on data distribution only, not knowing the class information [5]. That is, the similarity's adaptation to local data distribution is the main contributor in producing a data dependent kernel/similarity. In addition, this kernel does not need explicit learning, unlike the other three categories mentioned above.

This meeting facilitates an exchange of recent works and discussion around some of the fundamental questions/issues on this topic.

This document summarizes potential future research discussed in this Shonan meeting. It is organized into four sections:

(1) Potential future research

(2) Presentations provided in the meeting

(3) List of participants

(4) Meeting schedule

1 Potential future research

A number of interesting potential future research in relation to data dependent dissimilarity has been suggested during and after the Shonan meeting. This section summarises these potentials.

1.1 Using LID to estimate density ratio

Michael Houle

As part of the final-day discussions I presented an example of how an estimator of density could be derived using the Local Intrinsic Dimensionality (LID) model. Here, I give a slightly-improved overview of the derivation — for more background, please refer to the presentation slides titled "Local Intrinsic Dimensionality: An Extreme-Value-Theoretic Foundation for Similarity Applications" accompanying this report, or to the publications listed in the references at the end of the slides.

Often, estimation of density employs some measure of the volume of small balls with respect to the data domain — usually with respect to either the representational dimension, or the global intrinsic dimensionality. However, doing so presupposes that the data lies on a single manifold of some fixed dimension. In contexts (such as mixture models of local distributions of differing dimensionalities), measuring volume in this way can favor some of these distributions over others.

Particularly in contexts in which data dependent similarity metrics are to be employed, as an alternative to the estimation of density ratios with respect to the global distribution, one could instead estimate the ratios of densities with respect to the collection of distance distributions induced by the points of the data set.

Given two points $\mathbf{a}$ and $\mathbf{b}$ separated by a sufficiently small distance $d(\mathbf{a}, \mathbf{b}) = r$, let $A(r)$ and $B(r)$ be the probability measure associated with the ball of radius r centered at $\mathbf{a}$ and $\mathbf{b}$, respectively. Without a notion of global distribution, the density ratios in the vicinities of $\mathbf{a}$ and $\mathbf{b}$ can be defined as

$$\rho_{\mathbf{a},\mathbf{b}}(r) := \frac{A'(r)}{B'(r)} = \frac{A(r)\mathrm{ID}_A(r)}{B(r)\mathrm{ID}_B(r)} .$$

Here, we assume that the cumulative distribution functions A and B are both continuously differentiable over the range $r \in (0, c)$, thereby allowing us to use the relationship

$$\mathrm{ID}_F(r) = \frac{rF'(r)}{F(r)}$$

for $F \equiv A$ or $F \equiv B$.

This formulation of density ratio measures the ratio of the rates of expansion of probability measure, one rate viewed from $\mathbf{a}$ and the other from $\mathbf{b}$, if the distance r between $\mathbf{a}$ and $\mathbf{b}$ were to be allowed to vary. It can be regarded as a measure of the asymmetry of the distance measure d when adjusted for local probability measure and for local intrinsic dimensionality, or equivalently, the local discriminability of the original distance measure.

It should be noted that this ratio can break down in a theoretical sense if r is allowed to tend to zero. To see this, it suffices to consider the case where $A(r) \propto r^{m_{\mathbf{a}}}$ and $B(r) \propto r^{m_{\mathbf{b}}}$:

$$\lim_{r \to 0^+} \rho_{\mathbf{a},\mathbf{b}}(r) = \begin{cases} 0 & \text{if } m_{\mathbf{a}} > m_{\mathbf{b}} \\ 1 & \text{if } m_{\mathbf{a}} = m_{\mathbf{b}} \\ \infty & \text{if } m_{\mathbf{a}} < m_{\mathbf{b}} \end{cases} .$$

In a practical sense, for fixed $d(\mathbf{a}, \mathbf{b}) = r > 0$, we can approximate the density ratio by approximating the indiscriminabilities $\mathrm{ID}_A(r)$ and $\mathrm{ID}_B(r)$ by the local intrinsic dimensionalities ID_A^* and ID_B^*, respectively, where

$$\mathrm{ID}_F^* := \lim_{r \to 0^+} \mathrm{ID}_F(r)$$

for $F \equiv A$ or $F \equiv B$. ID_A^* and ID_B^* can then be estimated using existing techniques. This gives us the approximation

$$\rho_{\mathbf{a},\mathbf{b}}(r) \approx \frac{\mathrm{ID}_A^*}{\mathrm{ID}_B^*} \cdot \frac{A(r)}{B(r)} ,$$

where the ratio $A(r)/B(r)$ can be estimated by the ratio of the number of points within the balls of radius r centered at $\mathbf{a}$ and at $\mathbf{b}$. For those situations where these numbers of points are not available, or when the numbers are too low, the ID representation formula can help, as follows.

The ID representation states that as r and w tend to zero 'nicely' (that is, with r/w and w/r both bounded), we have that

$$\frac{F(r)}{F(w)} \to \left(\frac{r}{w}\right)^{\mathrm{ID}_F^*} .$$

By applying this twice, with w chosen as the distances within which A and B achieve probability $p > 0$ — that is, where $A(\delta_{\mathbf{a}}(p)) = p$ and $B(\delta_{\mathbf{b}}(p)) = p$, we obtain

$$\frac{A(r)}{B(r)} \approx \frac{A(\delta_{\mathbf{a}}(p))}{B(\delta_{\mathbf{b}}(p))} \left(\frac{r}{\delta_{\mathbf{a}}(p)}\right)^{\mathrm{ID}_A^*} \left(\frac{r}{\delta_{\mathbf{b}}(p)}\right)^{-\mathrm{ID}_B^*} = r^{\mathrm{ID}_A^* - \mathrm{ID}_B^*} \cdot \frac{(\delta_{\mathbf{b}}(p))^{\mathrm{ID}_B^*}}{(\delta_{\mathbf{a}}(p))^{\mathrm{ID}_A^*}} .$$

The distances $\delta_{\mathbf{a}}(p)$ and $\delta_{\mathbf{b}}(p)$ can then chosen as the respective k-NN distances of $\mathbf{a}$ and $\mathbf{b}$, so as to approximate the choice $p = k/n$, where n is the number of points in the data set.

Putting it all together, we get the following approximation for the density ratio:

$$\rho_{\mathbf{a},\mathbf{b}}(r) \approx r^{\mathrm{ID}_A^* - \mathrm{ID}_B^*} \cdot \frac{\mathrm{ID}_A^* \cdot (\delta_{\mathbf{b}}(k/n))^{\mathrm{ID}_B^*}}{\mathrm{ID}_B^* \cdot (\delta_{\mathbf{a}}(k/n))^{\mathrm{ID}_A^*}} .$$

As a final observation, when designing this estimator, we could have used choices of w at any convenient distance from $\mathbf{a}$ or $\mathbf{b}$. For this reason, the above estimator should be regarded only as an illustrative example. Many estimators are possible — in particular, formulations can be derived that use all distances within the k-NN sets of $\mathbf{a}$ and $\mathbf{b}$. It's an interesting question as to how effectively such estimators of density and density ratios may perform in practice. For the overall estimate to be stable, the estimates of the terms appearing in the exponents — ID_A^* and ID_B^* — must themselves be stable. This can possibly be achieved by blending the estimates over the respective neighborhoods of $\mathbf{a}$ and $\mathbf{b}$.

1.2 Alternative mass-based similarity which takes class distribution into account

Jaakko Peltonen

In many problem domains it is expected that not all statistical properties of the data variation will be relevant to analysts; for example part of the variation may be due to noise, known artifacts or distortions of the measurement process, or due to properties of the underlying phenomenon that are known, trivial, or otherwise uninteresting for analysis. Such uninteresting, noise or nuisance variation should not affect similarity or distance metrics that aim to solve exploratory or predictive tasks. Known data annotation of classes, ontologies, or constraints can help disambiguate which variation is relevant versus non-relevant, or more generally score how relevant each variation should be to the metric. In general the relevance of variation along variables, subspaces, of features is a local phenomenon. Such relevance needs to be combined with other desired properties of a metric such as adaptivity to data density.

In previous work [11] we have designed class annotation based local Riemannian metrics that are topology preserving and take local changes in class distributions into account. However, such metrics were not directly designed to be adaptive to data density, were not directly designed for robust estimation with finite data sets in high dimensionalities, and were designed as distance metrics instead of similarity measures. In new research, it is possible to research the connection between such metrics and estimation of similarity measures including relationship to mass-based similarities, and algorithms for efficient estimation of similarities that take into account both local changes of data density and local changes in relevance of data variation.

1.3 Relationship between dimensionality reduction and hubness

Miloš Radovanović, Michel Verleysen, John Lee, Frank-Michael Schleif

An important class of currently used performance measures for dimensionality reduction (DR) methods rely on the reasonable assumption that it is important for a DR method to preserve local neighborhoods from the original data in the reduced representational space. Apparently, this assumption has potential for strong interaction with the known property of hubness, with respect to models built from data, to unevenly distribute among the data points the responsibility for the errors that a model makes. Concretely for DR and neighborhood preservation, this leads to the hypothesis that hubness may cause the cost associated to "misplacing" a hub point in the reduced space to be significantly higher than that of non-hub points.

In the first phase, we will attempt to validate the above hypothesis by measuring neighborhood preservation errors associated with each data point, and correlating them with data point hubness. In case enough evidence is accumulated to support strong correlation, phase two will apply the obtained insights to modifying existing DR methods, at first focusing on the stochastic neighbor embedding (SNE) family, with the goal of producing DR methods that provide the correct amount of "special treatment" to hub points in order to improve neighborhood preservation. Along this line a strategy could be to modify the original cost function of SNE methods by adding an additional cost term or by changing the probability function in the divergence measure (again) such that potential hub points accumulating a higher weight in the divergence calculation than normal points. One may also impose an underlying topology in the two dimensional representation by constraining the positions of hubs and allowing more flexibility for other points in the low dimensional representation.

The proposed research not only has the potential to produce more effective DR methods in the short term, but also to provide better understanding of the underlying intricacies of high-dimensional spaces and mechanisms that can affect the DR process, leading to novel classes of methods and exciting new research directions in the long run.

1.4 Other issues related to Dimensionality reduction

John Lee, Michel Verleysen

1.4.1 Task-dependent dimensionality reduction

Dimensionality reduction (DR) is often cast within a completely unsupervised framework, similarly to clustering. When restricted to linear dimensionality reduction, some attempts have been made towards supervised methods, like projection pursuit to some extent and linear discriminant analysis. As to non-linear DR, most methods remain totally unsupervised and their typical purpose is exploratory data analysis in the broad sense. In this context, though, one might wonder whether the users have some intent or goal in mind and would like to particularize DR to suit their objectives. Typically this can cover the use of available class labels that are usually not involved in DR, except maybe in visualization or quality assessment. Another use case is meta-parameter browsing. Most DR methods have meta-parameters that impact their general behavior, like the scale or size of the considered neighborhoods, the trend to favor either false neighbors or missing ones, etc. Yet another aspect is the users' final intent. Do they use DR for itself and, if not, what is their true, underlying goal? Is it clustering, classification? For all these reasons, it might be useful to have DR methods where:

- some interactivity with user is possible
- metric learning can be integrated

Interactivity allows the users to test meta-parameters value and to assess visually or quantitatively the results. Optimization techniques should be adapted to enable seamless transitions between the subsequent problems that such use case implies. As to metric learning, DR is often considering default metric like the Euclidean distance. Task dependent metrics would be an improvement. A practical example is mixed data, where continuous, discrete, ordinal and categorical features are considered together, raising the question of their respective weight in the metric.

1.4.2 Dimensionality reduction based on data dependent similarity

Dimensionality reduction often relies on default choices for the metric, like the Euclidean distance. A few examples have tried to innovate with respect to this habit, like Isomap, where a graph of K-ary neighborhoods is used to approximate geodesic distances along the underlying manifold with shortest paths through the graph. To some extent, other spectral DR methods like Laplacian eigenmaps and variants implicitly use commute-time distances in K-ary neighborhood graphs. Similarities used in stochastic neighbor embedding are also a way of implicitly defining a data-dependent metric. Euclidean distances are wrapped up in softmax similarities whose normalization brings invariances. Moreover, the bandwidth in the exponential allow for adaptation to local data density. Multi-scale similarities push that idea even further (multiple invariances, multiple bandwidths). A perspective that is not yet explored is to use anisotropic distances, like when switching from Euclidean to (local) Mahalanobis. Although it thought to be useful, such data-dependent metric involves many additional parameters that need to be adjusted carefully.

1.5 A generic proof of kernel characteristic of Isolation Kernel

Kai Ming Ting and Ye Zhu

Isolation Kernel, implemented using iForest, has been shown to have the following characteristic empirically:

> 'Two points in sparse region are more similar than two points of equal inter-point distance in dense region.' [5] (A formal description of this characteristic is provided in the paper.)

A recent work has used nearest neighbour partitioning mechanism (aka Voronoi diagram) to induce Isolation Kernel [6]. Although a proof is provided for the above-mentioned characteristic of Isolation Kernel, the proof relies on this particular implementation [6].

The characteristic of Isolation Kernel is independent of its partitioning mechanism, as long as the partitions created satisfy the requirement that 'large partitions in sparse regions and small partitions in dense regions'. Thus, a generic proof of the characteristic, independent of the partitioning mechanism, is sought.

Such a generic proof will provide (i) a better understanding of the kernel's behaviour; (ii) a guidance to further development of Isolation Kernel in terms of designing different partitioning mechanisms and potential deviations from the stated characteristic; and (iii) a connection to mass-based similarities [7,8] which have the same similarity characteristic (and can be implemented using the same partitioning mechanism) but derived from a different formulation.

1.6 Feature Map of Isolation Kernel

Jaakko Peltonen and Kai Ming Ting

What are the equivalent features (provided intrinsically) from Isolation Kernel?

There are many uses of the features derived this way. Examples are: (a) Random sampling of these features have been employed to reduce the computational cost of employing the full set of features derived from a data independent kernel. Features derived from Isolation Kernel can be similarly applied. (b) It facilitates the application of random projection.

1.7 Is Concentration of Measure an issue in practice?

Kai Ming Ting, John Lee, Michel Verleysen, Takashi Washio, Ye Zhu

Is concentration of measure [12,13] an issue in practice?

Email discussion was conducted on this question after the meeting. Some have the view that 'real' datasets appear to have low intrinsic dimensions and sufficient structure—this has kept the concentration effect at bay in practice. Some has the view that the current 'real' datasets have been influenced by the data collection methods thus far. This can change in the future. The study of this effect in high dimensional problems shall not be ignored.

1.8 Other issues related to data-induced similarities

Kai Ming Ting

The following issues were raised in the meeting:

(a) What are the base kernels for different isolation partitioning mechanisms for Isolation Kernel?

The base kernel refers to the data independent kernel that Isolation Kernel approximates under uniform density distribution. Using the approximation derived by Leo Breiman, Isolation Kernel implemented using iForest has been shown to approximate to Laplacian Kernel under uniform density distribution [5].

An implementation using the Voronoi diagram [6] has a different base kernel; so as any other implementations of Isolation Kernel. The knowledge of the base kernel facilitates practitioners in choosing the 'prior' for a particular problem at hand.

(b) How to make Isolation Kernel more adaptive to different aspects of data characteristics, tailored for a specific task.

The existing Isolation Kernels [5,6] are unsupervised, i.e., the data dependency is solely based on data distribution. While Isolation Kernel has been shown to perform better than distance metric learning (which utilizes the class information) in SVM classifiers [5], it is possible that utilizing additional information in the data may further improve the task-specific outcome.

In another perspective, existing Isolation Kernels do not have explicit learning. It would be interesting to investigate incorporating an optimization process to further enhance Isolation Kernel for the task at hand.

(c) Data-induced similarity of different characteristics

Isolation Kernel is one type of data-induced similarity which has a specific kernel characteristic that is akin to one aspect of human-judged similarity (as described in Section 1.5). There are other aspects of human-judged similarity [9,10]. It would be interesting to examine other types of data-induced similarity which have these characteristics; and whether they have practical advantages over data independent similarity.

(d) Similarity for mixed data types

Data-induced similarity is focused on numeric attributes only. Extending its ability to handle categorical attributes and mixed attribute types will benefit both the scientific community as well as industry.

References

[1] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. Proceedings of the 15th International Conference on Neural Information Processing Systems, 521-528, 2002.

[2] Mehmet Gönen and Ethem Alpaydin. Multiple kernel learning algorithms. Journal Machine Learning Research, 12:2211-2268, 2011.

[3] Shun-Ichi Amari and Si Wu. Improving support vector machine classifiers by modifying kernel functions. Neural Network, 12(6):783-789, 1999.

[4] Leo Breiman. Some infinity theory for predictor ensembles. Technical Report 577. Statistics Dept. UCB, 2000.

[5] Ting, K.M., Zhu, Y. and Zhou, Z-H. (2018) Isolation Kernel and Its effect on SVM. Proceedings of 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2329-2337.

[6] Qin, X-Y., Ting, K.M., Zhu, Y. and Lee, C.S. Nearest-Neighbour-Induced Isolation Similarity and Its Impact on Density-Based Clustering. Appear in AAAI 2019.

[7] Ting, K.M., Zhu, Y., Carman, M., Zhu, Y. and Zhou Z-H. (2016) Overcoming key weaknesses of distancebased neighbourhood methods using a data dependent dissimilarity measure. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1205-1214.

[8] Aryal, S., Ting, K.M., Haffari, G. and Washio, T. (2014) m_p-Dissimilarity: A Data Dependent Dissimilarity Measure. Proceedings of IEEE ICDM. 707-712.

[9] A. Tversky (1977) Features of similarity. Psychological Review, 84(4):327-352.

[10] Krumhansl, C. L. 1978. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review 85(5):445-463.

[11] Peltonen, J., Klami, A., and Kaski, S. (2004) Improved Learning of Riemannian Metrics for Exploratory Data Analysis. Neural Networks, 17:1087-1100.

[12] Michel Talagrand. (1996) A new look at independence. The Annals of Probability, 24(1):1-34.

[13] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. (1999) When is nearest neighbor meaningful? ICDT 99, 217-235.

2 Presentations provided in the meeting

A total of 12 presentations were planned and made on the first three days of the meeting; and 3 short presentations were added on the last day. The presentation slides are shared among participants using DropBox.

(i) Data-Induced Similarities

- Mass-based dissimilarity: Kai Ming Ting & Takashi Washio

  Abstract: The use of distance metrics such as the Euclidean or Manhattan distance for nearest neighbour algorithms allows for interpretation as a geometric model, and it has been widely assumed that the metric axioms are a necessary condition for many data mining tasks. We show that this assumption can in fact be an impediment to producing effective models. We propose to use mass-based dissimilarity, which employs estimates of the probability mass to measure dissimilarity, to replace the distance metric. This substitution effectively converts nearest neighbour (NN) algorithms into Lowest Probability Mass Neighbour (LMN) algorithms. Both types of algorithms employ exactly the same algorithmic procedures, except for the substitution of the dissimilarity measure. We show that LMN algorithms overcome key shortcomings of NN algorithms in classification and clustering tasks. Unlike existing generalised data independent metrics (e.g., quasi-metric, meta-metric, semi-metric, peri-metric) and data dependent metrics, the proposed mass-based dissimilarity is unique because its self-dissimilarity is data dependent and non-constant.

- Learning with non-metric proximities: Frank-Michael Schleif

  Abstract: Efficient learning of a data analysis task strongly depends on the data representation. Most methods rely on (symmetric) similarity or dissimilarity representations by means of metric inner products or distances, providing easy access to powerful mathematical formalisms like kernel or branch-and-bound approaches. Similarities and dissimilarities are however often naturally obtained by non-metric proximity measures which can not easily be handled by classical learning algorithms. In the last years major efforts have been undertaken to provide approaches which can either directly be used for such data or to make standard methods available for these type of data. The presentation provides a comprehensive overview for the field of learning with non-metric proximities. First we introduce the formalism used in non-metric spaces and motivate specific treatments for non-metric proximity data. Secondly we provide a systematization of the various approaches. For a few approaches we discuss complexity issues and generalization properties. We also address the problem of large scale proximity learning which is often overlooked in this context and of major importance to make the method relevant in practice. The discussed algorithms and concepts are in general applicable for proximity based clustering, one-class classification, classification, regression or embedding tasks. Various applications show the relevance of the discussed approaches, which provide a generic framework for multiple input formats. The goal of the presentation

is to give an overview about recent developments in this domain, covering in particular principled approaches as concerns learning in indefinite spaces and its mathematical foundations and extensions to large scale problems.

- Mutual Reachability Distances: Ricardo J. G. B. Campello

  Abstract: In this talk I elaborate on how a family of mutual reachability distances can transform the original data space in a way that distances can be locally stretched in a data-dependent way, and how this type of transformation plays a fundamental role in recent effective and efficient density-based data mining algorithms. In particular, I discuss their application to density-based clustering, outlier detection, and semi-supervised classification.

- Some thoughts on learning theory for metric learning: Yiming Ying

  Abstract: Metric learning has attracted a large amount of interest. Despite many algorithms that have been proposed, there is little work on the statistical foundation to explain the empirical successes behind such methods. For instance, a fundamental question is how to characterize the generalization ability and classification performance of such algorithms in terms of how well they perform on new (test) data when trained on given historical data.

  In this talk I will present our efforts in this challenging research direction. Firstly, we show that generalization analysis of supervised metric learning reduces to the estimation of Rademacher average over "sums-of-i.i.d." sample-blocks. Then, we derive generalization bounds for metric/similarity learning with different matrix-norm regularizers by estimating their Rademacher complexities. Our analysis using U-statistics and Rademacher complexity indicates that sparse similarity learning with L1-norm regularization can lead to significantly better bounds than those with Frobenius-norm regularization. Secondly, we address the links between similarity learning and the classification performance of the resulting classifier. We show that the generalization error of the resulting classifier can be bounded by the generalization bound of similarity learning. This shows that a good generalization of the learned similarity function guarantees a good classification of the resulting classifier.

- Isolation Kernel and its effect on SVM: Kai Ming Ting

  Abstract: This presentation reports a recent data dependent kernel that is derived directly from data. Data dependent kernel has been an outstanding issue for about two decades which hampered the development of kernel-based methods. We introduce Isolation Kernel which is solely dependent on data distribution, requiring neither class information nor explicit learning to be a classifier. In contrast, existing data dependent kernels rely heavily on class information and explicit learning to produce a classifier. We show that Isolation Kernel approximates well to a data independent kernel function called Laplacian kernel under uniform density distribution. With this revelation, Isolation Kernel can be viewed as a data dependent kernel that adapts a data independent kernel to the structure of a dataset.

We also provide a reason why the proposed new data dependent kernel enables SVM (which employs a kernel through other means) to improve its predictive accuracy. The key differences between Random Forest kernel and Isolation Kernel are discussed to examine the reasons why the latter is a more successful tree-based kernel.

(ii) Rescaling and Dimensionality reduction

- CDF-based rescaling—An effective way to deal with inhomogeneous density datasets: Ye Zhu

  Abstract: Density-based clustering algorithms find clusters in regions of high density which are separated by regions of low density. The clusters are typically identified by grouping points which are above a global density threshold. They are able to find clusters of arbitrary sizes and shapes while effectively separating noise. Despite its advantage over other types of clustering, it is well-known that most density-based algorithms face the same challenge of finding clusters with varied densities.

  In this talk, I present a principled density-ratio approach that enables a density-based clustering algorithm to identify clusters with varied densities. Density-ratio estimate the ratio between the density of a given point and the average density of its local neighbourhood. Existing density-based methods could use density-ratio estimation to find clusters as regions of local high densities, which are separated by regions of local low densities.

  Instead of reconditioning an existing density-based algorithm with the density-ratio estimator, I provide three CDF-based rescaling methods as a pre-processing step to rescale a given dataset, then existing density-based algorithm can be applied unaltered to the rescaled dataset to perform density-ratio based clustering. These three rescaling methods can be treated as density equalisation w.r.t. a density estimator with a certain bandwidth such that different clusters shows similar densities after rescaling. Therefore, a single density threshold can be used to identify all clusters that would otherwise impossible had the same algorithm been applied to the unscaled dataset.

- Multiscale stochastic neighbor embedding: John A. Lee

  Abstract: Nonlinear dimensionality reduction (DR) is also known as manifold learning and consists in determining a low-dimensional representation of high-dimensional data, which is somehow faithful to their initial salient features, like underlying manifolds, clusters, outliers, or distribution in general. The consensual proxy for the task of DR is that dissimilar data items should be represented far apart, whereas similar items should lie close to each other. This proxy is implemented as such for the quality assessment of DR methods and results. In practice DR QA evaluates the correct average preservation of K-ary neighbourhoods around all data points, for various values of size K. Over the history of DR, this proxy has also been instantiated in various way to design DR methods, like the preservation of data variance in principal component analysis (PCA) or the preservation

of dot products in classical multidimensional scaling (CMDS) or distance preservation in stress-based MDS. More recently, improved results have been obtained with similarity preservation, where distances are wrapped into similarities with specific invariance properties. In particular, the similarities used in stochastic neighbour embeddings and variants are interpreted as probabilities for two points to be neighbours and the normalization in their softmax ratio makes them invariant to squared distance shifts. This property is decisive to discrepancy in distance concentration between the HD data space and the LD representation. Another property is related to the cost function used to carry out DR, which is analogous to an energy functional in an N-body placement problem (masses and springs). For some cost functions, the gradient mimicks spring plasticity (overly elongated springs deform and do no longer oppose an attractive force). Plasticity allows 'tearing' manifold, leading to improved neighbourhood preservation on average. Eventually, a last feature of DR methods is how they consider the issue of scale. Linear DR tends to preserve large neighbourhoods in priority, whereas nonlinear modern methods tend to preserve smaller, local neighbourhoods. This is sometimes specified explicitly, like in SNE and variants, where the user provides a scale parameter called 'perplexity'. We present a method where scale must no longer be specified. Similarities are then averages over multiple scales, covering all possible neighbourhood sizes. Experimental results show that this approach yields improved results while also dispensing the user with the selection of a preferred scale.

- Dimensionality reduction—Targeted Projection Pursuit: Maia Angelova

  Abstract: Big data come with high dimensions. Reducing dimensionality without losing essential information, hidden in the data is number one priority of dimension-reduction techniques. Another priority is the ability to visualise data in a format accessible by users who are not necessarily specialists in the field.

  Targeted Projection Pursuit (TPP) [Faith et al Bioinformatics 2006], is a dimension reduction machine learning method that allows the exploration of previously clustered data on the two-dimensional screen.

  The theory behind the TPP method is discussed and some applications for re-clustering and classifications are given. These applications include gene expression data for leukemia cancers [Faith et al, Bioinformatics 2006] and data for the usage of telecare devices for North East of England [Angelova et al IEEE Access 2018] will be presented. The performance of the method is demonstrated compared to other dimensionality reduction techniques. Further directions for development to include time series and stochastic data are discussed.

  References:

  J Faith, R Mintram and M Angelova. Targeted Projection Pursuit for Gene Expres-sion Data. Bioinformatics, 22, No 21, pp 26672673 (2006), doi:10.1093/bioinformatics/btl463.

  Maia Angelova, Jeremy Ellman, Helen Gibson, Paul Oman, Suthar-

shan Rajasegarar, Ye Zhu. User Activity Pattern Analysis in Telecare Data. IEEE Access, Issue Date: DECEMBER 2018, 6, Issue:1, 33306-33317, DOI: 10.1109/ACCESS.2018.2847294.

- Similarities and Learning with Random Projections: Ata Kaban

  Abstract: Random projection (RP) is a simple, computationally efficient and theoretically well grounded dimensionality reduction technique. This talk highlights some aspects of its relationship with the notion of similarity in Euclidean data spaces. (1) We show that, with high probability, RP preserves dot products and their sign to an extent that depends on the cosine similarity of the vectors. (2) Furthermore, we show that, for halfspace learning, the use of RP yields new generalisation bounds in terms of the expectation of a function of cosine similarities between the classifier and points of the input space. Such bounds hold without any assumptions beyond i.i.d. sampling of the data, both in the case of learning from RP-ed data, as well as for learning from the original data, and have close links with margin distribution bounds. (3) Moreover, RP is not confined to linear models. An analogous application of the ideas to the nearest neighbour classifier reveals geometric characteristics that explain the statistical difficulty or easiness of a problem for nearest neighbour - this turns out to be a notion of intrinsic dimension of the input space. (4) A further example we give is learning a Mahalanobis metric within a generic classifier, which is known to have a sample complexity that strongly depends on the data dimension in general. Here a RP-based analysis highlights the Frobenius norm of the similarity matrix as a measure of problem difficulty, and this recovers a known finding in a special case. (5) Finally, we note that learning a Fisher linear discriminant (FLD) classifier may be interpreted as learning the sign of a Mahalanobis dot-product similarity, and we show how this can be done in small sample conditions by means of an ensemble of FLDs that each receive an independent RP-ed version of the data. We give theoretical guarantees as well as state of the art empirical results for this approach, including an analysis of the number of base learners necessary for the similarity matrix of the ensemble to reach arbitrarily close in spectral norm, with high probability, from that of the infinite ensemble.

- Nonlinear Dimensionality Reduction with Missing Values: Michel Verleysen

  Abstract: Dimensionality reduction (DR) aims at faithfully and meaningfully representing high-dimensional data into a low-dimensional (LD) space. Recently developed neighbor embedding DR methods lead to outstanding performances, thanks to their ability to foil the curse of dimensionality. Unfortunately, they cannot be directly employed on incomplete data sets, which become ubiquitous in machine learning. Discarding samples with missing features prevents their LD coordinates computation and deteriorates the complete samples treatment. Common missing data imputation schemes are not appropriate in the nonlinear DR context either. Indeed, even if they model the data distribution in the feature space, they can at best

enable the application of a DR scheme on the expected data set. In practice, one would instead like to obtain the LD embedding with the closest cost function value on average with respect to the complete data case. As state-of-the-art DR techniques are nonlinear, the latter embedding results from minimizing the expected cost function on the incomplete database, not from considering the expected data set. This paper addresses these limitations by developing a general methodology for nonlinear DR with missing data, being directly applicable with any DR scheme optimizing some criterion. In order to model the feature dependencies, a high-dimensional extension of Gaussian mixture models is first fitted on the incomplete data set. It is afterward employed under the multiple imputation paradigm to obtain a single relevant LD embedding, minimizing the cost function expectation. Extensive experiments demonstrate the superiority of the suggested framework over alternative approaches.

Reference: Cyril de Bodt, Dounia Mulders, Michel Verleysen and John Aldo Lee. Nonlinear Dimensionality Reduction With Missing Data Using Parametric Multiple Imputations. IEEE Transactions on Neural Networks and Learning Systems, published online 27 August 2018, DOI: 10.1109/TNNLS.2018.2861891

(iii) Related issues

- Hubness and Data Dependent Dissimilarity Measures: Miloš Radovanović

  Abstract: Hubness – 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 summarizes the knowledge and recent research on hubness, making the connections with data dependent dissimilarity measures (DDDMs), and is organized in three parts: (1) Origins of hubness, which discusses 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 related to DDDMs, which presents some notable effects of (anti-)hubs on techniques for machine learning, data mining and information retrieval, identifies two different approaches to handling hubs adopted by researchers – through fighting or embracing their existence – and reviews techniques and applications belonging to the two groups, with particular focus on their relationship with DDDMs; and (3) Discussion, which initiates dialogue about open problems and areas with significant opportunities for research on connections between hubness and DDDMs.

- Local Intrinsic Dimensionality: An Extreme-Value-Theoretical Foundation for Similarity Applications: Michael E. Houle

  Abstract: Researchers have long considered the analysis of similarity applications in terms of the intrinsic dimensionality (ID) of the

data. This presentation is concerned with a generalization of a discrete measure of ID, the expansion dimension, to the case of smooth functions in general, and distance distributions in particular. A local model of the ID of smooth functions, LID, is first proposed and then explained within the well-established statistical framework of extreme value theory (EVT). Moreover, it is shown that under appropriate smoothness conditions, the cumulative distribution function of a distance distribution can be completely characterized by an equivalent notion of data discriminability. As the local ID model makes no assumptions on the nature of the function (or distribution) other than continuous differentiability, its extreme generality makes it ideally suited for the non-parametric or unsupervised learning tasks that often arise in similarity applications. The LID model is then extended to a multivariate form that can potentially account for the contributions of different distributional components towards the intrinsic dimensionality of the entire feature set, or equivalently towards the discriminability of distance measures defined in terms of these feature combinations. Formulas are established for the effect on LID under summation, product, composition, and convolution operations on smooth functions in general, and cumulative distribution functions in particular. For some of these operations, the dimensional or discriminability characteristics of the result are also shown to depend on a form of distributional support. Finally, a theoretical relationship is established between the LID model and the classical correlation dimension.

(iv) Additional presentations provided on the last day

- David Gao
- An Information Retrieval Approach to Visualization of High-dimensional Data, and Learning Metrics from Annotation and Interaction: Jaakko Peltonen

  Abstract: In this talk we discuss how neighbor embedding can be formalized as an information retrieval task, its performance can be measured by information retrieval measures, and the embeddings can be directly optimized for the information retrieval task, by the Neighbor Retrieval Visualizer (NeRV) method family. Several variants of the method family are presented. We also discuss how a topology-preserving Riemannian local information-theoretic metric, the Learning Metric, can be derived from data annotation, and how it can be used in neighbor embedding. We further discuss how a metric can be learned iteratively from annotation gathered from visual interaction with human experts.
- Hiroshi Motoda

3 List of Participants

Kai Ming Ting	Federation University	kaiming.ting@federation.edu.au
Takashi Washio	Osaka University	washio@ar.sanken.osaka-u.ac.jp
Ata Kaban	University of Birmingham	ata.x.kaban@gmail.com
Frank-Michael Schleif	University of Applied Sciences Würzburg-Schweinfurt	fmschleif@googlemail.com
Hiroshi Motoda	Osaka University	motoda@ar.sanken.osaka-u.ac.jp
Jaakko Peltonen	University of Tampere	Jaakko.Peltonen@uta.fi
James Bailey	University of Melbourne	baileyj@unimelb.edu.au
John Aldo Lee	Université Catholique de Louvain	john.lee@uclouvain.be
Masayuki Numao	Osaka University	numao@sanken.osaka-u.ac.jp
Michael Houle	National Institute of Informatics	meh@nii.ac.jp
Michel Verleysen	Université Catholique de Louvain	michel.verleysen@uclouvain.be
Yiming Ying	State University of New York at Albany	yying@albany.edu
Arthur Zimek	University of Southern Denmark	zimek@imada.sdu.dk
David Gao	Federation University	d.gao@federation.edu.au
Ye Zhu	Deakin University	ye.zhu@deakin.edu.au
De-Chuan Zhan	Nanjing University	zhandc@lamda.nju.edu.cn
Maia Angelova	Deakin University	maia.a@deakin.edu.au
Tomoyuki Higuchi	Institute of Statistical Mathematics	higuchi@ism.ac.jp
Ricardo Campello	University of Newcastle	ricardo.campello@newcastle.edu.au
Miloš Radovanović	University of Novi Sad	radacha@dmi.uns.ac.rs
Peer Kröger	Ludwig-Maximilians-Universität München	kroeger@dbs.ifi.lmu.de
Kaoru Yoshida	Sony Computer Science Laboratories	kaoru@csl.sony.co.jp

Meeting Schedule

		AM1		AM2		PM1		PM2	PM3
	8:45-9:00	9:00-10.30		10:45-12:15		1:00-2:30		2:45-4:15	4:15-5:45
15-Oct	Intro	K.M. Ting & T. Washio		Frank-Micheal		Richardo Campello		Yiming Ying	Kai Ming Ting
16-Oct		Ye Zhu		John Lee		Maia Angelova		Ata Kaban	Michel Verleysen
17-Oct		Milos Radovanovic		Micheal Houle		Excursion + Banquet			
18-Oct		Discussion Session		Discussion Session					

Each presenter has one hour for presentation and 30 minutes for discussion

6