様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２０　日現在

機関番号：　６２６１５
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７５３

研究課題名（和文）Practical and Effective Data Mining Via Local Intrinsic Dimensional Modeling

研究課題名（英文）Practical and Effective Data Mining Via Local Intrinsic Dimensional Modeling

研究代表者

Ｍｉｃｈａｅｌ　Ｅ．Ｈｏｕｌｅ（Houle, Michael E.）

国立情報学研究所・大学共同利用機関等の部局等・客員教授

研究者番号：９０３９９２７０

交付決定額（研究期間全体）：（直接経費）　　12,400,000円

研究成果の概要（和文）：ビッグデータの時代には、データボリュームが非常に大きくなったわけで、従来のデータ処理アプリケーションを使用できなくなってくるのである。データマイニング、機械学習、マルチメディアなどの分野での類似検索については、ソフトウェア実装の効率と有効性は、データの類似性の測定とデータオブジェクトの特徴に依存する。特徴の数（データ次元の数）が高い場合、多くの無関係な特徴の属性によって類似性の測定にエラーの原因になり、関連する特徴を圧倒し得る。このプロジェクトの目標は、内在的な次元性の局所的な変化を利用できる技術を開発し、ビッグデータの類似性アプリケーションにおける次元性の問題に取り組むことである。

研究成果の概要（英文）：In the era of Big Data, data volumes have become so enormous and so complex as to preclude processing using traditional applications. For similarity search and retrieval, as well as many other fundamental operations in such areas as data mining, machine learning, multimedia, recommendation systems, and bioinformatics, the efficiency and effectiveness of software implementations depends crucially on the interplay between measures of data similarity and the features (or attributes) by which data objects are represented. When the number of features (the data dimensionality) is high, the errors introduced into similarity measurements by the many irrelevant feature attributes can overwhelm the contributions of the relevant features. The overall goal of this project is to tackle the problem of the curse of dimensionality in similarity applications for big data, by developing practical unsupervised techniques that recognize and take advantage of local variations in intrinsic dimensionality.

研究分野：　Data mining, indexing, similarity search

キーワード：　高次元空間　極値理論　データマイニング

１．研究開始当初の背景

In the era of Big Data, data volumes have become so enormous and so complex as to preclude processing using traditional applications. For similarity search and retrieval, as well as many other fundamental operations in such areas as data mining, machine learning, multimedia, recommendation systems, and bioinformatics, the efficiency and effectiveness of software implementations depends crucially on the interplay between measures of data similarity and the features (or attributes) by which data objects are represented. For such applications (which we refer here to as similarity 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 number of features (the data dimensionality) is high, the errors introduced into similarity measurements by the many irrelevant feature attributes can completely overwhelm the contributions of the relevant features. As the dimensionality of data increases, the discriminative ability of similarity measures diminishes to the point where techniques that depend on search lose their effectiveness.

The negative effects of the curse of dimensionality on similarity applications are well-known and well-documented, and can manifest themselves differently from discipline to discipline. For example, in multimedia applications, the feature sets ('visual vocabularies') currently being sought for large image corpuses or video archives can have millions of members ('visual words'). These huge feature set sizes can lead to severe problems for clustering, classification, and any other forms of analysis that depend on content-based similarity measures. Another example is data mining, where data analysis often require the identification of relatively small clusters of mutually-similar data objects (nuggets). Identification of nuggets usually entails the generation of similarity queries to determine the candidates for cluster membership. High dimensionality prevents traditional clustering methods from identifying nuggets.

Although high-dimensional data sets are very often difficult to process, it is not always the case. Data sets of high dimensionality can often be searched and analyzed relatively easily, provided that the so-called intrinsic dimensionality is low. The intrinsic dimensionality of a dataset can loosely be regarded as the number of latent variables needed to describe a data set: this could be the number of 'relevant' or 'non-redundant' features, or the degrees of freedom of a distribution that accurately models the data.

At the beginning of this project, low intrinsic dimensionality had already made an impact in practice, in guiding the design and performance analysis of certain similarity search methods. The theory and practice had been limited to global approaches, where the intrinsic dimensionality is expressed as a single value for an entire dataset. However, this does not fit well with the user's experience of the data, where interest may be focused on a specific query result or a specific data cluster. Just as data is often modeled as a mixture of underlying distributions, data sets have regions of varying intrinsic dimensionality – and these local variations should be modeled theoretically and exploited in practice.

Researchers had also given attention to a measure of local variation within data sets: with respect to a collection of fixed-size neighborhoods based at each of the objects of a data set, the hubness of a point x is defined as the number of these neighborhoods that contain x. Hubness has been shown to be empirically correlated with both high intrinsic dimensionality, and the centrality of the object within the cluster that contains it – this phenomenon has been exploited in accelerating the performance of clustering algorithms. Our preliminary findings indicated that the hubness of points within data distributions can be fully quantified and explained under the new continuous ID model.

２．研究の目的

The overall goal of this project was to tackle the problem of the curse of dimensionality in similarity applications for big data, by developing practical unsupervised techniques that recognize and take advantage of local variations in intrinsic dimensionality. More precisely, the goals were:

(1) To advance the theory of intrinsic dimensionality so as to account for preexisting empirical research on the hubness of data, as well as the known effectiveness of shared neighbor information;
(2) To confirm the theoretical implications by means of a detailed empirical study of ID, hubness, and SNN in unsupervised similarity applications;
(3) To exploit this new knowledge of ID, hubness and SNN to develop more efficient and more effective solutions for unsupervised applications of data mining and multimedia, including data clustering, anomaly detection, feature selection, and variants of similarity search;
(4) To make these technical innovations available to researchers and practitioners through their integration into practical systems for search and analysis of data, as well as through publication in top-level international journals and conference proceedings.
(5) To create and promote a new interdisciplinary international research community, focused on the issues surrounding the unsupervised applications of ID within the areas of databases, data mining, and multimedia.

３．研究の方法

The project was a 3-year international collaboration among 5 main participants (the Principal Investigator plus 4 overseas Research Collaborators), each one a member of a different research institute or university. Collectively, the participants have expertise that spans the areas of computer science that are touched by the curse of dimensionality, including: theory, algorithms, and similarity applications for databases, data mining, and multimedia (search and indexing, clustering, outlier detection, feature selection). Each of the main team members was responsible for organizing contributions to the project from among their own research collaborators and students. In the first year, the concentration was to be on the empirical validation of the theory of local intrinsic dimensionality on real scenarios in databases and data mining (similarity search, clustering, outlier detection, etc.). In 2016 and 2017, the work was to extend to the use of ID to achieve original breakthroughs in more challenging settings, such as unsupervised feature selection, subspace clustering, and ensembles of distance measures. Successful outcomes were expected to have an immediate impact via integration into a well-known data mining development platform. Throughout the project, our goal was also to create and promote a new interdisciplinary international research community on the theme of dimensionality and scalability.

The research team members and their roles were as follows (current affiliations are shown):

- Vis. Prof. Michael E. Houle, National Institute of Informatics (NII), Japan, PI: theoretical computer science, algorithms, data mining, indexing, similarity applications, computational geometry. His main expertise is in the design and analysis of algorithms for similarity search, clustering, outlier detection and feature selection that exploit intrinsic dimensionality. The original local intrinsic dimension model was originally proposed by him.
- Assoc. Prof. Dr. Arthur Zimek, University of Southern Denmark (USD), Denmark, RC: data mining, clustering, outlier detection. Prof. Zimek is a leading authority in outlier detection, subspace clustering, and similarity search, having co-authored several prominent papers in these areas. He is also the principal author of the well-known ELKI data mining application development platform.
- Prof. Vincent Oria, New Jersey Institute of Technology (NJIT), USA, RC: databases, multimedia indexing. Prof. Oria has collaborated closely with the PI on applications of ID to adaptive similarity queries and subspace similarity search, sharing the supervision of 3 NJIT PhD students.
- Dr. Laurent Amsaleg, IRISA Rennes, France, RC: multimedia databases, multidimensional indexing, content-based retrieval, database security and privacy. Dr. Amsaleg is part of a research group (at IRISA Rennes) that has developed practical multimedia indexing systems capable of searching billions of images, represented by hundreds of thousands of features. In the lead-up to the project, he has collaborated with Houle on the development of estimators of local ID.
- Assoc. Prof. Miloš Radovanović,

University of Novi Sad, Serbia, RC: Prof. Radovanović is the foremost authority on the hubness phenomenon, as well as algorithms for data clustering which exploit it.
- Prof. James Bailey, University of Melbourne, Australia, RC: Prof. Bailey was invited to join the project after the first year, due to his leading reputation and expertise in data mining and machine learning. His group has had a particular interest in problems involving the modeling of data by subspaces – a highly promising area for the application of local intrinsic dimensional techniques.

Each of these main team members was responsible for organizing contributions to the project from among their own research collaborators and students.

４．研究成果

Of the goals laid out at the beginning of the project, we have fully met our objectives in with respect to all 5 research goals. The second goal, the confirmation of theoretical implications by means of a detailed empirical study of ID, was done instead in separate contexts – due to the explosive progress in local ID theory and the opportunity for application to problems of importance in indexing, data mining, and machine learning, we came to realize that it would be more efficient and appropriate to evaluate the theory context by context rather than in one comprehensive study.

The main discovery of the project is that the local ID model can explain, in a very useful and intuitive way, some of the important characteristics of learning and classification using deep neural networks. In particular, we have obtained the following results:

- Submission (and eventual acceptance) of a refereed international conference paper at the top-ranked ICML machine learning conference (to be presented in July 2018) [1]. This paper shows that local ID can be used to track the learning behavior of deep neural networks, to the extent of being able to detect when the learning process has stabilized. This detection allows the learning process to be terminated, thereby avoiding overfitting.
- Acceptance of a refereed international conference paper at the top-ranked ICLR machine learning conference (presented on 2018/5/1) [3]. Our paper deals with the problem of detecting adversarial examples, which are carefully crafted instances that can mislead deep neural networks (DNNs) to make errors during prediction. We show that adversarial examples can be characterized in terms of the unusually high local intrinsic dimensionality (LID) within the data space surrounding them, according to the model of LID developed in this project. Our paper was one of only 23 papers (out of 935) accepted for full oral presentation (for an acceptance rate of 2.5%).
- Presentation of a refereed international conference paper at WIFS 2017, in which a theoretical analysis is given showing that the vulnerability of classification to adversarial perturbation increases as the local intrinsic dimensionality rises [4].

Other highlights of the research outcomes for this project include:

- Presentation of 3 refereed international conference papers (at SISAP 2017). Two of these papers together lay out the full theoretical

foundations of the LID model, its connection to similarity search and extreme value theory, and an extension to the multivariate setting [6,7]. The third paper showed the power of the LID model for the practical guiding the selection of local features suitable for the formation of high quality similarity graphs [5]. A related study has been accepted for publication in June 2018 [2].
- A refereed international journal publication which uses the theory of local intrinsic dimensionality to guide the heuristic termination of content-based similarity search processes involving multiple query objects [8].
- Publication of a refereed international journal paper in the Proceedings of the VLDB Endowment (PVLDB) [9]. This paper proposes an efficient method for the important yet difficult reverse k-nearest neighbor search problem. Our solution uses run-time optimization (including early termination) guided by an intrinsic dimensional testing criterion. The method significantly outperforms its competitors, particularly in that it can make use of

our existing estimators of local ID for autotuning its heuristic choices.
- One refereed international conference paper on the use of local ID to measure dependency in data [10]. The paper shows that the LID-based criterion can simultaneously identify multiple functional relationships in real data that conventional measures cannot handle.
- One refereed international journal paper on the topic of similarity search within a projected subspace, where the features identifying the subspace are supplied at query time [11]. Here as well, dimensional testing is employed so as to accelerate performance.
- In a top refereed international journal, an empirical study of unsupervised outlier detection measures, datasets and methods. Although this work is not specific to local ID, this experimental framework was developed for the evaluation of ID-based outlier detection methods that are likely to be developed in the future.
- Two refereed international full papers (one conference and one journal) on the topic of flexible aggregate similarity search [13,15]. State-of-the-art solutions are given using dimensional testing, which employs estimates of ID at runtime to control the tradeoff of execution time versus query accuracy.
- Publication of a refereed international conference full paper at the top-tier ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) [14]. This paper proposed, analyzed, and evaluated several efficient estimators for local intrinsic dimensionality (ID) based on Extreme Value Theory (EVT), and lays the practical foundation for the applications of ID that were investigated in this project.

In addition to smaller workshops on local ID held at NII and at IRISA-Rennes, team members led the organization of two conferences at which local intrinsic dimensional analysis was introduced and promoted:

- Successful organization in Tokyo of the 9th International Conference on Similarity Search and Applications (SISAP 2016), which we are cultivating as a venue for promoting recent advances in the area of local intrinsic dimensionality.
- Successful organization of the second NII Shonan Meeting on Dimensionality and Scalability.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　15　件）

[1] Xingjun Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S.-T. Xia, S. Wijewickrema, and J. Bailey, “Dimensionality-Driven Learning with Noisy Labels”, 35th International Conference on Machine Learning (ICML), July 2018, 10 pages, DOI なし, 査読あり.

[2] B. Bratić, M. E. Houle, V. Kurbalija, V. Oria, M. Radovanović, “NN-Descent on High-Dimensional Data”, 8th International Conference on Web Intelligence, Mining and Semantics (WIMS), June 2018, 8 pages, DOI なし, 査読あり.

[3] Xingjun Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, J. Bailey, “Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality”, 6th International Conference on Learning Representations (ICLR), April 2018, pp. 1−15, DOI なし, 査読あり.

[4] L. Amsaleg, J. Bailey, D. Barbe, S. M. Erfani, M. E. Houle, X. V. Nguyen, and M. Radovanović, “The Vulnerability of Learning to Adversarial Perturbation Increases with Intrinsic Dimensionality”, 9th IEEE Workshop on Information Forensics and Security (WIFS), December 2017, pp. 1−6, DOI 10.1109/WIFS.2017.8267651, 査読あり.

[5] M. E. Houle, V. Oria, and A. M. Wali, “Improving k-NN Graph Accuracy using Local Intrinsic Dimensionality”, 10th International Conference on Similarity Search and Applications (SISAP), October 2017, pp. 110−124, DOI 10.1007/978-3-319-68474-1_8, 査読あり.

[6] M. E. Houle, “Local Intrinsic Dimensionality II: Multivariate Analysis and Distributional Support”, 10th International Conference on Similarity Search and Applications (SISAP), October 2017, pp. 80−95, DOI 10.1007/978-3-319-68474-1_6, 査

読あり.
[7] M. E. Houle, “Local Intrinsic Dimensionality I: An Extreme-Value-Theoretic Foundation for Similarity Applications”, 10th International Conference on Similarity Search and Applications (SISAP), October 2017, pp. 64–79, DOI 10.1007/978-3-319-68474-1_5, 査読あり.
[8] M. E. Houle, Xiguo Ma, V. Oria, and J. Sun, “Query Expansion for Content-Based Similarity Search Using Local and Global Features”, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 13(3), August 2017, pp. 25:1–25:3, DOI 10.1145/3063595, 査読あり.
[9] G. Casanova, E. Englmeier, M. E. Houle, P. Kröger, M. Nett, A. Zimek, “Dimensional Testing for Reverse k-Nearest Neighbor Search”, Proceedings of the VLDB Endowment (PVLDB), 10(7), March 2017, pp. 769–780, DOI 10.14788/3067421.3067426, 査読あり.
[10] S. Romano, O. Chelly, X. V. Nguyen, J. Bailey, and M. E. Houle, “Measuring Dependency via Intrinsic Dimensionality”, 23rd International Conference on Pattern Recognition (ICPR), December 2016, pp. 1207–1212, DOI 10.1109/ICPR.2016.7899801, 査読あり.
[11] M. E. Houle, Xiguo Ma, V. Oria, and J. Sun, “Efficient Similarity Search within User-Specified Projective Subspaces”, Information Systems, 59, July 2016, pp. 2–14, DOI 10.1016/j.is.2016.01.008, 査読あり.
[12] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, “On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study”, Data Mining and Knowledge Discovery (DAMI), 30(4), July 2016, pp. 891–927, DOI 10.1007/s10618-015-0444-8, 査読あり.
[13] M. E. Houle, Xiguo Ma, and V. Oria, “Effective and Efficient Algorithms for Flexible Aggregate Similarity Search in High Dimensional Spaces”, IEEE Transactions on Knowledge and Data Engineering (TKDE), 27(12), December 2015, pp. 3258–3273, DOI 10.1109/TKDE.2015.2475740, 査読あり.
[14] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, and M. Nett, “Estimating Local Intrinsic Dimensionality”, 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 2015, pp. 29–38, DOI 10.1145/2783258.2783405, 査読あり.
[15] M. E. Houle, Xiguo Ma, and V. Oria, “Flexible Aggregate Similarity Search in High-Dimensional Data Sets”, 8th International Conference on Similarity Search and Applications (SISAP), 9371, October 2015, pp. 15–28, DOI 10.1007/978-3-319-25087-8_2, 査読あり.

６．研究組織

(1)研究代表者

フール　マイケル　E.（HOULE, Michael E.）

国立情報学研究所・大学共同利用機関等の部局等・客員教授

研究者番号：90399270