氏名	Oussama Chelly	
学位(専攻分野)	博士(情報学)	
学位記番号	総研大甲第 1929 号	
学位授与の日付	平成29年3月24日	
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当	
学位論文題目	Intrinsic Dimensionality: from Estimation to Applications	
論文審査委員	主　査	教授　河原林　健一
		教授　宇野　毅明
		准教授　市瀬　龍太郎
		教授　佐藤　真一　国立情報学研究所
		客員教授　HOULE, Michael　国立情報学研究所
		教授　鷲尾　隆　大阪大学

(別紙様式 2)
(Separate Form 2)

論文の要旨
Summary (Abstract) of doctoral thesis contents

The thesis is primarily concerned with a new form of local, distance-based intrinsic dimensionality (LID) with ties to extreme value theory. Most commonly, the intrinsic dimensionality within a data region is expressed in terms of the number of dimensions (or basis vectors) required to describe a subspace (or manifold) that best approximates the data. Distance-based intrinsic dimensionality, on the other hand, takes a distributional view of data, in which the dimension is inferred from the distribution of distances of data objects to a local reference point. The submitted thesis is primarily concerned with the development of effective estimators of LID, and the application of LID estimation to feature selection.

The thesis consists of 8 chapters, the first 4 of which can be considered to be introductory in nature:

1. Chapter 1 provides the motivation for and history of the use of intrinsic dimensionality for such problems as dimensionality reduction, manifold learning, and feature selection. A distinction is made between global and local intrinsic dimensionality, and distance-based intrinsic dimensionality is identified as the topic of the thesis.
2. Chapter 2 comprehensively surveys the state of the art in intrinsic dimensional estimation. However, the descriptions for most of the estimation methods are brief, with only those methods most relevant to the thesis topic (expansion-based methods) being covered in detail.
3. Chapter 3 gives a brief and accessible overview of extreme value theory, with only the minimum details needed to support the discussions in Chapter 5. Attention is given to the limit distribution within the lower tail of the distribution of distances to a query point, under appropriate assumptions of smoothness (continuous differentiability).
4. Chapter 4 introduces the theory behind the local intrinsic dimensionality (LID) model that is adopted in this thesis. Starting from the extreme value theory (EVT) result that the upper tail of a smooth distribution follows a Generalized Pareto Distribution, a connection between LID and EVT is stated.
5. Chapter 5, the first original contribution of the thesis, is an investigation of estimation methods for local intrinsic dimensionality (LID), in which established distance-based statistical estimation techniques (both general and EVT) are considered. Here, the candidate shows that the maximum likelihood estimator for LID is equivalent to the Hill estimator from EVT, and that this estimator is the most effective in practice from among a collection of competing

state-of-the-art estimators. A more general class of estimators based on regularly varying (RV) functions is also shown to relate to existing estimators, depending on the parameter choices: for one choice, RV is equivalent to the Generalized Expansion Dimension (GED); for another, it converges to the Hill estimator. Finally, an analysis of the variance of RV estimation is provided, which shows that the variance is minimized by GED.

6. Although applications such as outlier detection or classification generally require information drawn from small neighborhoods, local intrinsic dimensional estimation generally requires neighborhoods of at least 100 points. Chapter 6 addresses this problem by proposing the ALID estimator, which uses auxiliary distances (distances among pairs of points within the neighborhood of a test point) as well as the usual neighbor distances. A theoretical analysis is given that shows that the number of auxiliary distances available is highest when the intrinsic dimensionality is low. The experimental results provided show that ALID has significantly better convergence properties than estimators based on direct distances alone, and allows for estimation over significantly smaller neighborhoods, but only when the intrinsic dimensionality is not too high (which is the case in many practical settings).
7. Chapter 7 presents an application of LID estimation to feature selection. An existing approach based on local variance (Laplacian Score) is adapted to work on LID values instead – this is motivated by an interpretation of LID in terms of the indiscriminability of the distance measure when using a given feature. Two heuristic algorithms are proposed (“univariate” and “multivariate”) and analyzed theoretically. Performance guarantees are given through an argument based on the submodularity of the cost function. The experimental results show first that the state of the art competitors for local feature selection do not compete well even against uniform random feature selection, and second that the proposed LID-based algorithms generally match or exceed the performance of random selection.
8. Chapter 8 presents a conclusion and a discussion of possible future extensions of this work. These extensions include the possibility of developing estimators for second-order intrinsic dimensionality, and other potential applications of ID estimation.

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨
Summary of the results of the doctoral thesis screening

The examiners were generally satisfied with the thesis final defense presentation, and with the performance of the candidate in responding to the questions posed during the examination. All comments and suggestions from the preliminary defense were addressed to the satisfaction of the examiners.

From the thesis and the defense, the examiners were able to identify the following main original contributions in the thesis:

1. (Chapter 5) Results on LID estimation: (a) the demonstration that the Hill estimator from extreme value theory applies to LID, and is the most effective in practice; and (b) the characterization of the family of RV estimators, and the analysis of minimum variation of RV estimation. This result has been published in a top international conference (KDD 2015) and accepted for publication in an international journal (DAMI), and is likely to find applications in practice.
2. (Chapter 6) The introduction of auxiliary distances (distances among pairs of points within the neighborhood of a test point) is shown to lead to estimators with better convergence properties than estimators based on direct distances alone. This is quite likely to have an impact in practice, as it addresses a specific difficulty encountered in many data mining and machine learning applications.
3. (Chapter 7) When applied to feature selection, the use of LID as a criterion leads to competitive performance relative to the state-of-the-art. The committee notes that LID and its competitors often struggle to outperform random feature selection, which may lower the practical impact of the proposed selection techniques.

The examiners note that the candidate is the lead author of one refereed international journal paper accepted for publication, and one refereed international conference paper presented, both of which are included in the thesis (Chapter 5).

The examiners also note that two other manuscripts drawn from the thesis material (Chapter 6 and Chapter 7) have also been submitted for publication at refereed international conferences. One of these is pending, and the other has been rejected and is awaiting resubmission.

The thesis content and contributions have been judged satisfactory upon evaluation

(別紙様式 3)
(Separate Form 3)

(from the presentation and the thesis proper), and the examiners thus unanimously recommended that the candidate be awarded the degree of 博士（情報学）.