Data Complexity Analysis: Linkage between Context and Solution in Classification

Tin Kam Ho

Bell Labs, Alcatel-Lucent
tkh@research.bell-labs.com

Abstract. For a classification problem that is implicitly represented by a training data set, analysis of data complexity provides a linkage between context and solution. Instead of directly optimizing classification accuracy by tuning the learning algorithms, one may seek changes in the data sources and feature transformations to simplify the data geometry. Simplified class geometry benefits learning in a way common to many methods. We review some early results in data complexity analysis, compare these to recent advances in manifold learning, and suggest directions for further research.

1 Introduction

Challenges from practical problems represented by publicly shared data sets have contributed much to the wide participation and diverse advances in statistical pattern recognition. The data sets, and the baseline accuracies of standard algorithms on such, provided a reference frame for evaluating the effects and differential advantages of many new methods and their variants. Researchers of classification algorithms benefited from being isolated from the often tedious and resource-intensive stage of preparing training data. However, an undesirable side effect is that they were also isolated from the context of the recognition problem, driven to concentrate on obtaining incremental progresses in accuracies using minor tweaks on the learning algorithms. Such a narrow view of a pattern recognition problem resulted in many uninteresting experiments that do not lead to useful conclusions.

Data complexity analysis attempts to address this dilema. In developing a way for characterizing a data set by its geometrical and topological complexity, we study whether, and to what extent, the class boundaries are learnable. Detailed quantification of learnability provides operational guidance on the expected accuracies of a family of algorithms that share common geometrical behavior. It allows for the return of attention to many issues surrounding the compilation of the training data, and provides a chance to relate the classification task to its original context. In the cases where accuracy is constrained by the geometrical properties inherent in the data, one is prompted to look for improvements

upstream, that is, improvements in the data collection, feature extraction, or feature transformation stages that will lead to reduction in data complexity.

Early explorations in data complexity analysis have followed an empirical approach, and resulted in several discoveries: (1) a large collection of problems are shown to span a continuum in several measures of geometrical complexity, with no obvious gaps and clusters; (2) apparently there exist a small number of complexity measures that provide independent characterization of different important aspects relevant to classification; (3) there are identifiable domains of dominant competence for a set of standard classifiers; and (4) measurable complexity may change due to feature selection and feature transformation. In this article we review these early progresses, discuss related efforts in newer approaches, and suggest some future directions.

2 Parameterization of Data Complexity

Complexity Classes and Complexity Scales

The concern in data complexity analysis is on characterizing each instance of the classification problem. This is different from the worst case analysis that is common for combinatorial optimization problems and algorithms.

In classification, early discussions on data complexity focused on the split between linearly separable problems and those that are not. The discovery of the perceptron's inability to learn linearly nonseparable classes led to a long silence in activities in neural networks, until newer structures and learning algorithms were found to overcome the obstacle. In the meantime, efforts were made to stretch linear classifiers to adapt to complex class boundaries. Piecewise linear classifiers, polynomial discriminants, low-dimensional feature projections and high-dimensional feature transformations, nearest-neighbor classifiers, and decision trees were among the popular pursuits. Optimism built up with breakthroughs in multi-layer perceptron training, support vector machines, and ensemble learning methods. Many arguments were made on the generality of the ability of large-capacity learning methods on arbitrary problems. Accuracy limits observed in empirical studies are believed to be setbacks that can be addressed by careful control of overfitting.

Yet we believe that such practical limits deserve a deeper investigation.

In the pursuit of multiple classifier systems that may benefit from the merits of all the existing learning algorithms, and the pursuit of generative algorithms like random decision forests that may create as many classifiers as needed, it graduately became clear that some problems are intrinsically very difficult to learn under the lack of a complete sample along with class-labeling. It is difficult for a small-sample representation of the problem to contain complete knowledge of the classes, especially in high-dimensional spaces. There is a need for a bet-

ter understanding of the class geometry to provide guidance to the creation of classifier components. Data complexity analysis was motivated by this need.

Continuum of Complexity Distribution

Early studies [14] revealed that a collection of many benchmarking data sets span a nearly continuous range in several complexity measures. The complexity measures include known descriptors of class geometry, such as Fisher's discriminant ratio, ratio of between-class and within-class distances, departure from convexity and smoothness of boundaries, fragmentation of the support of the classes, and length of the class boundary. In addition, the error rates of several simple classifiers, such as nearest-neighbor classifiers, linear classifiers induced by linear programming minimizing the sum of error distances, can also be considered useful parameters of data complexity.

For a standard problem, e.g., discrimination of two classes represented by two Gaussians of different means, it is understandable that changes in the distributional parameters may lead to differences in classification difficulty. Say, with the variances fixed, increasing distance between the means promotes separability, and such changes can be continuous.

The surprise in the discovery is that the observed continuum is spanned by a large but almost arbitrary collection of benchmarking data sets. This gives raise to a doubt that the common practice of arguing about relative merits of learning algorithms on a very limited collection of problem instances can be misleading. Simple counts of the number of times an algorithm "wins", or even numerical measures of the best, worst, average accuracies over the entire set, have little relevance to the algorithm's performance on a future problem of unknown complexity. Better characterization of the merits of learning algorithms should involve a careful, balanced selection of problem instances to cover differences in data complexity in different aspects.

Intrinsic Dimensionality of Complexity Distribution

Another observation from the early experiments is that the distribution of the problem instances in the complexity space is far from uniform. The distribution concentrates on a narrow band across the space, but there are outliers. Correlation between some measures explains the narrow distribution in the projection to certain dimensions. For example, it is observed that the error rate of a nearest-neighbor classifier is highly correlated with the fraction of points located on the class boundary. This can be explained by a close examination of how a boundary point contributes to nearest-neighbor classification error. On the other hand, fragmentation and sphericity of the classes have very low correlation with measures of separation distances and class convexity [15]. In one experiment, principal component analysis shows that six linear combinations of the measures would explain 94% of the variance [14].

If the behavior of other classifiers is used to parameterize data complexity, one can easily adopt too many measures for the analysis to be practical [23]. Examples include the size of a decision tree grown to fit the training data, number of nearest neighbors needed after condensation, or the type and parameters of kernels useful in support vector machines. Arguably many of these gauge the data complexity in some sense. However, even after removing those that do not provide good characterization for problems with known complexity, there are too many possibilities. Clustering the measures offers a solution [23], but a more interesting approach will be to uncover those measures that directly describe the essential geometrical and topological characteristics that determine the behavior of the more sophisticated classifiers. We shall return to this subject when we discuss the interplay between the learning algorithms and geometrical complexity.

Bayes Error and Uncertainty in Complexity Estimates

In classification, the analysis is complicated by the fact that while the application context can be reasonably specified, the full difficulty of a problem can only be understood through a sample data set. A consequence of such indirectness is that a fair analysis of the problem's complexity is affected by the sampling rate and sampling distribution. Only in very rare cases, a generative model is available that can produce an infinite amount of representative data [13].

In classical statistics, a classification problem's difficulty is quantified by the Bayes error, which can be calculated if the probability density function of each class is known. The Bayes error gives a lower bound on a classifier's asymptotic error rate. The lower bound is achievable by parametric classifiers suited to the distributional characteristics in some simple problems (e.g. two Gaussians). Some nonparametric methods like nearest-neighbor classifiers are also shown to be able to closely approach the Bayes error with infinite training samples [8].

When logistic regression is used for classification, the uncertainty of class prediction can be described by the confidence interval on the log-odds prediction. For other classifiers, a large literature exists on relating sample sizes to classification error [9][24], and relating generalization error to sample size and classifier capacity [7][27][28].

Given problems of the same Bayes error, differences in their geometrical complexity can cause large differences in classification accuracies because of the simple geometrical models used by many standard classifiers. However, the characterization of data geometry is subject to uncertainty due to small sample sizes. This issue is especially severe if the data are given as dissimilarity representations rather than feature vectors [10]. In lack of additional data, one may have to invoke assumptions on the problem domain, or general heuristics such as the compactness hypothesis [10].

3 Relating Classifier Behavior to Data Complexity

Domain of Classifier Competence

In an analysis of several classifiers' relative merits in the data complexity space, concentrations are observed in the distribution of problem instances where some classifiers have dominant advantage. Most notable is that the domain of dominant competence of the linear classifier and that of the nearest-neighbor classifier occupy opposite ends of the complexity distribution when projected to a measure of boundary length and a measure of boundary nonlinearity [19]. Similarly, one can specify the region where a certain classifier has advantages [20], or compare the competence regions of two similar classifiers [12]. One useful result from these studies is that several complexity measures, including the boundary length, boundary nonlinearity, and the ratio of intra-inter class nearest-neighbor distances, are the most relevant metrics for discriminating between competence domains of different classifiers [16].

However, while interesting observations like these have been made, the study is far from conclusive. On close examination of the competence domains, one can see substantial spread and overlap. Given the similarities of the classifiers under study, it is conceivable that their competence domains are intrinsically ambiguous. It is unclear to what extent the ambiguity can be alleviated by using additional complexity measures. The correlation between the complexity measures complicates the issue. Furthermore, one has to consider the uncertainty in the complexity estimates for each problem due to small sample sizes.

Another difficulty is that in our study using empirical data sets, certain regions in the complexity space are not well covered. The uneven coverage makes it difficult to obtain a comprehensive mapping from the complexity measures to classifier preferences. In this discrimination task, large uncertainty in the prediction can occur due to small sample sizes (few problem instances) in certain regions. To address this difficulty, a systematic scheme to select or synthesize sample problems will be needed.

Changes in Complexity due to Feature Selection

Feature selection has long been an important theme in classification. Often, the goal is to reduce the demand on run time and computer memory, and to alleviate the effect of the curse of dimensionality, in other words, to avoid overfitting irrelevant noise contained in the data.

A study on selecting subsets of features using an LPSVM (Linear Programming Support Vector Machine) formulation and a Forward Feature Selection procedure shows that such selection can change the data complexity substantially [21]. A note of caution is that the changes may occur only in the *apparent* complexity, i.e., the complexity estimate calculated from the particular data set. If the sample

size is too small for the problem, the changes may not be a reliable indicator of changes in the problem's intrinsic difficulty.

4 Experimental Design for Classifier Evaluation

Experiences from the early studies point to the need for further efforts in selecting appropriate complexity measures and designing a better classifier evaluation strategy. This can be compared to a conventional computational experiment, where critical issues include a proper parameterization of the input domain and a good way to generate samples to cover the domain.

In pattern recognition, this amounts to simulating classification problems arising from different contexts. The goal is to provide a comprehensive coverage of (or a close approximation to) all possible problems that may arise in a practical application. Going back to the application context, one may seek modifications of the problem formulation, so that the classification task can be matched to one with the lowest data complexity. A few approaches can be considered in designing the evaluation strategy.

Synthesis of Classification Problems with Target Complexity

In this approach, synthesizing a classification problem involves creating a suitable distribution of the training data points in the feature space and their class labeling. Ideally, the synthesis procedure provides an even cover of the chosen complexity measure. In [18] an interesting attempt is made on covering the range of a boundary length measure. The experiment first generates n data points that follow a uniform random distribution in an m-dimensional space. A minimum spanning tree (MST) is calculated for each realization. Different choices in designating some edges to be class-crossing result in data sets of different complexity according to this measure. The reference numbers for n, m may come from a real-world problem. To compare classifier accuracies on the synthetic data with those on real-world data, the same number of MST edges are made to be class-crossing. Therefore, the synthetic data are at the same apparent complexity with the real-world data under this measure. Observation of the classifier accuracies shows that this method provides a pessimistic estimate of achievable accuracy. A possible cause is that the real-world problems have a more compact geometry than the uniform distribution in the synthetic data, so that they are actually easier. To close this discrepancy, additional complexity measures are needed to guide the production of synthetic data towards similar compactness.

Synthesis of Problems with References to Natural Processes

A more plausible approach to create synthetic data is to follow certain well known models of stochastic spatial processes, e.g., the Neyman-Scott process, that represent certain natural processes. Using such models allows for formal inferences on the resulting data complexity. Furthermore, uncertainty in the complexity estimates can be evaluated in a principled way.

Systematic Degradation of Real-World Problems

One may also produce new classification problems by systematically degrading existing ones. An advantage of this approach is that the process starts with a realistic data geometry. The drawback is that the reference geometry also introduces a bias, which needs to be removed by sufficient coverage using a larger variety of real-world problems. This could be very difficult for high dimensional spaces.

A version of this method is to take a real-world problem that has a parameterized domain model. Perturbing the model parameters leads to different realizations of the problem with different complexity. For example, in [13] a document image defect model is used to create increasing difficult instances of a character recognition problem. Other variations include simulations of imperfect sampling conditions or class labeling errors.

5 Simplification of Class Geometry

A constructive procedure to improve classification performance is to seek to reduce data complexity by changing a problem's formulation, introducing clever feature extractors, and mapping the classification task to an easier space by feature selection and transformation. This is an age-old theme in pattern recognition. But we argue for more than just a return of attention to this theme.

Instead of relying on intuition and heuristics, we argue for a way to formalize this process and set goals for systematic optimization. Description of data complexity gives an operational definition of learnability, against which systematic optimization can be driven. An extreme example is that one may seek to extract enough features so that the classification task is reduced to be linearly separable in the lowest dimensionality. Another example is to include clever normalization procedures and observe problem-specific invariances that can compress the within-class scatter. In many engineering contexts this practice has been followed implicitly or explicitly. Ideally, more automation can be introduced into the process by linking the complexity measures to controllable processes in the upstream.

Manifold Learning and Dimensionality Reduction

Recent advances in the methods for manifold learning provide some interesting directions. Several methods have been proposed to fit localized linear models to the data manifold. The local models are embedded into a global model along the intrinsic dimensions. In the process the data cloud is unfolded to a smooth and flat surface. The emphasis of many manifold learning techniques is on obtaining a useful transformation that highlights the intrinsic dimensions, yields better clustering, and helps visualization. Some transformations can be applied to unseen data from the same source (out-of-sample extensions) [5].

Most manifold learning methods do not give explicit characterization of the data geometry. This is deferred to a subsequent clustering step. Also, little is said about the existence of holes, class fragmentations, etc. Recent efforts in topological data analysis may address these concerns [6]. Other interesting attempts include modeling entire data sets and their similarities by Grassmann manifolds [26], and applying the embedding methods to tensor representations [11].

A *supervised* manifold learning procedure highlights the geometry of the class boundary and promotes linearization of the separating surface. A pioneering work in this area is [25], where a supervised locally linear embedding (LLE) method is proposed. By applying LLE to a distance matrix where between-class distances are exaggerated, the authors show that superior classification accuracy can be achieved especially on high-dimensional data where the class structures contain curved low-dimensional manifolds. An symptom of such data is a high contrast between the global dimensionality and local intrinsic dimensionality. More recent work extends this to tensor representations [17].

In many of these studies, the merits of the entire approach, including both the feature transformation step and the classification and clustering step, are evaluated jointly as if the steps are parts of one complete method. This incurs a risk that the classification method chosen may not be the optimal companion to the feature transformation step. With data complexity analysis, an intermediate goal can be made explicit. One may seek a manifold learning/mapping technique that reduces the complexity of the data geometry, and let the next step be guided by a known match of data complexity to classifier choices.

6 Application Examples

The limit of achievable recognition accuracy remained a subject for the experts for a long time, until recently, a few high-profile projects brought this to public attention. A few examples are as follows.

A clever way to exploit the limits of mechanical pattern recognition is the CAPTCHA system [1]. The system challenges a user to enter the text that appears as an image that is distorted from its regular shape, in an arbitrary font, corrupted with noise and overprinting. It is believed that the recognition task is beyond machine reach, so that a user who enters the text correctly is trusted to be a living human. This is also known as a reverse Turing test, or human interactive proofs [3].

The Netflix challenge [22] was open to the public since 2006. The challenge is a public contest on improvements to the video rental company's system for predicting movie rating for its users based on the user's previous ratings on other movies. The project offers a prize to the first team with an algorithm that can achieve a 10% improvement on the test set over the company's in-house algorithm. While progress has been made since the beginning of the contest, at

the time of this paper's writing, it remains unclear whether the project's goal is intrinsically achievable.

Another example is the list of "Human Intelligence Tasks" offered at the Amazon web service "Mechanical Turk" [2]. Many of these are small pattern recognition tasks that machines do not perform well, for example, identifying and tagging objects in an image sequence. Do these task represent the limit of power in automatic pattern recognition? Or do we expect that, someday in the future, a fair portion of these can be fully automated by a smarter classifier technology?

7 Conclusions

We reviewed the development of data complexity analysis over the last decade, and summarized the experiences from early explorations. We argued that data complexity provides an important focus for the optimization of the formulation and solution of a classification task. It provides a linkage between optimizations related to the problem context and those related to the learning algorithms.

Our discussion has focused primarily on the feature vector space representation of classification problems. We have not addressed challenges arising from alternative representations (e.g. similarity based representations, tensors, categorical variables, strings), data processing issues (unit and scale normalization, missing values, mislabeling), or multi-layered or meshed dependence structures. We also note that more can be explored on coupling data complexity analysis with active and sequential learning.

Acknowledgements

My collaborators on this endeavor have contributed many of the ideas summarized in this article: Mitra Basu, Ester Bernado-Mansilla, Richard Baumgartner, Martin Law, Erinija Pranckeviciene, Albert Orriols-Puig, and Nuria Macia. I would also like to thank Henry Baird, George Nagy, Anil Jain, Robert Duin, Josef Kittler, Ludmila Kuncheva, David Hand, Thomas Bengtsson, Peter Bickel, Bin Yu, Emina Soljanin, and participants of the Multiple Classifier Systems workshops for many stimulating discussions.

References

1. L. von Ahn, M. Blum, N.J. Hopper, J. Langford, CAPTCHA: Telling Humans and Computers Apart. In *Advances in Cryptology, Eurocrypt'03*, volume 2656 of Lecture Notes in Computer Science, 2003, 294-311. Also see: http://www.captcha.net/.
2. Amazon, Mechanical Turk, 2005, http://www.mturk.com/mturk/welcome.
3. H.S. Baird, Complex Image Recognition and Web Security, in [4], 287-298.
4. M. Basu, T.K. Ho (eds.), *Data Complexity in Pattern Recognition*, London: Springer-Verlag, 2006.
5. Y. Bengio, J-F. Paiement, P. Vincent, Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering, *NIPS 2003*, 177-184.

6. G. Carlsson, Topology and Data, Dept of Math., Stanford Univ., preprint, Aug 10, 2008. http://comptop.stanford.edu/preprints/topologyAndData.pdf
7. V. Cherkassky, Y. Ma, Data Complexity, Margin-Based Learning, and Popper's Philosophy of Inductive Learning, in [4], 91-114.
8. T.M. Cover, P.E. Hart, Nearest Neighbor Pattern Classification, *IEEE Trans. on Inf. Theory*, **13** 1967, 21-27.
9. L. Devroy, Automatic Pattern Recognition: A Study of the Probability of Error, *IEEE Trans. on Pat. Anal. and Mach. Intell.*, **10**, 4, 1988, 530-543.
10. R.P.W. Duin, E. Pekalska, Object Representation, Sample Size, and Data Set Complexity, in [4], 25-58.
11. X. He, D. Cai, P. Niyogi, Tensor Subspace Analysis, *NIPS 2005*.
12. T.K. Ho, A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors, *Pattern Analysis and Applications*, **5**, 2002, 102-112.
13. T.K. Ho, H.S. Baird, Large-Scale Simulation Studies in Image Pattern Recognition, *IEEE Trans. on Pat. Anal. and Mach. Intell.*, **19**, 10, October 1997, 1067-1079.
14. T.K. Ho, M. Basu, Complexity Measures of Supervised Classification Problems, *IEEE Trans. on Pat. Anal. and Mach. Intell.*, **24**, 3, March 2002, 289-300.
15. T.K. Ho, M. Basu, M.H.C. Law, Measures of Geometrical Complexity in Classification Problems, in [4], 3-23.
16. T.K. Ho, E.B. Mansilla, Classifier Domains of Competence in Data Complexity Space, in [4], 135-152.
17. X. Li, S. Lin, S. Yan, D. Xu, Discriminant Locally Linear Embedding with Higher-Order Tensor Data, *IEEE Trans. on Sys., Man, and Cyb.*, Part B: Cyb., **38**, 2, 2008, 342-352.
18. N. Macia, E.B. Mansilla, A. Orriols-Puig, Preliminary Approach on Synthetic Data Sets Generation Based on Class Separability Measure, *Proc. of the 19th Int'l Conf. on Pat. Recog.*, Tampa, U.S.A., Dec 7-11, 2008.
19. E.B. Mansilla, T.K. Ho, On Classifier Domains of Competence, *Proc. of the 17th Int'l Conf. on Pat. Recog.*, Cambridge, U.K., Aug 22-26, 2004, Vol. 1, 136-139.
20. E.B. Mansilla, T.K. Ho, Domain of Competence of XCS Classifier System in Complexity Measurement Space, *IEEE Trans. on Evol. Comp.*, **9**, 1, February 2005, 82-104.
21. E. Pranckeviciene, T.K. Ho, R. Somorjai, Class Separability in Spaces Reduced By Feature Selection, *Proc. of the 18th Int'l Conf. on Pat. Recog.*, Hong Kong, China, Aug 20-24, 2006, Vol. 2.
22. Netflix Prize, 2006, http://www.netflixprize.com/.
23. S. Raudys, Measures of Data and Classifier Complexity and the Training Sample Size, in [4], 59-68.
24. S. Raudys, A.K. Jain, Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners, *IEEE Trans. on Pat. Anal. and Mach. Intell.*, **13**, 3, 1991, 252-264.
25. D. de Ridder, O. Kouropteva, O. Okun, M. Pietikainen, R.P.W. Duin, Supervised Locally Linear Embedding, *Art. Neural Net. and Neural Inf. Proc.* (Proc. ICANN/ICONIP 2003), Lect, no. 2714, Berlin: Springer-Verlag, 2003, 333-341.
26. A. Srivastava, A Bayesian Approach to Geometric Subspace Estimation, *IEEE Trans. Sig. Proc.*, **48**, 5, May 2000, 1390-1400.
27. V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Berlin: Springer-Verlag, 1982.
28. V. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.