Fused Feature Representation Discovery for High-Dimensional and Sparse Data

Jun Suzuki and **Masaaki Nagata**
NTT Communication Science Laboratories, NTT Corp.
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
{suzuki.jun, nagata.masaaki}@lab.ntt.co.jp

Abstract

The automatic discovery of a significant low-dimensional feature representation from a given data set is a fundamental problem in machine learning. This paper focuses specifically on the development of the feature representation discovery methods appropriate for high-dimensional and sparse data. We formulate our feature representation discovery problem as a variant of the semi-supervised learning problem, namely, as an optimization problem over unsupervised data whose objective is evaluating the impact of each feature with respect to modeling a target task according to the initial model constructed by using supervised data. The most notable characteristic of our method is that it offers a feasible processing speed even if the numbers of data and features are both in the millions or even billions, and successfully provides a significantly small number of feature sets, *i.e.*, fewer than 10, that can also offer improved performance compared with those obtained with the original feature sets. We demonstrate the effectiveness of our method in experiments consisting of two well-studied natural language processing tasks.

Introduction

The automatic discovery of a significant low-dimensional feature representation from a given data set, which we refer to as '*feature representation discovery*', has been a long-standing goal of machine learning research. Many different feature representation discovery methods have already been developed, and their usefulness has been demonstrated in many areas of real data analysis, including text, speech, image, and signal data processing. For example, PCA, SVD, ICA, and modern variants (Van Der Maaten and Hinton 2008) are typical feature representation discovery methods that seek a low-dimensional representation via feature/data matrix decomposition. One example that directly seeks sparse representation is sparse coding (Zhang et al. 2011). Standard clustering methods have also been utilized to capture reduced representations, and have mainly been applied to tasks with discrete feature spaces (Turian, Ratinov, and Bengio 2010). More recently, several new ideas have been proposed including 'word-codebook' (Kuksa and Qi 2010), which tries to capture an abstract of words as a low-dimensional real valued vector, and deep learning (Hinton 2007), which seeks a representation that captures higher level, abstract features of the given data using a multi-layer neural network.

Let X be an $M \times N$ feature/data matrix, where M is the number of features, and N is the number of data points. In this paper, we focus on a situation where the task at hand is extremely large, *i.e.*, M and N are both in the millions or even billions. Today, we often encounter such large-scale problems, especially in text processing and bio-informatics, since the automated collection of (unsupervised) data is rapidly increasing, and many fine-grained features must be incorporated to improve task performance. Thus, developing specialized algorithms suitable for dealing with large-scale problems is an important research topic in machine learning.

The difficulty presented by large problems mainly originates in the computational cost of the solver algorithms. For example, polynomial time algorithms are obviously infeasible when the numbers of features and/or data points are in the millions. Another difficulty may derive from data sparsity, *i.e.*, more than 90% or even 99% of the elements in matrix X are zero, since, in general, large-scale problems tend to be very sparse problems. Under this condition, for example, PCA, ICA and their variants are essentially useless since most of the data points are orthogonal to each other. Additionally, obtained new feature representations should satisfy the condition that they can be calculated without incurring a large additional computational cost.

Against this background, the goal of this paper is to provide a feasible and appropriate method for tackling extremely large feature representation discovery problems. First, we formulate our feature representation discovery problem as a variant of a semi-supervised learning problem, namely, an optimization problem over unsupervised data whose objective is to evaluate the impact of each feature with respect to modeling a target task according to the initial model, which is constructed by using supervised data. Our method has the following three main characteristics; (1) it has the ability to handle a large number of data *i.e.*, the order of billions, since our method is designed to work in distributed computing environments, (2) it can work appropriately even if the original feature set is an infinite set, and (3) the resultant new feature representation generated by our method consists of only a very small number of features, *i.e.*,