Information Theoretic Clustering of Sparse Co-Occurrence Data

Inderjit S. Dhillon and Yuqiang Guan
Department of Computer Sciences
University of Texas
Austin, TX 78712-1188, USA
inderjit, yguan@cs.utexas.edu

UTCS Technical Report #TR-03-39

September 12, 2003

Abstract

A novel approach to clustering co-occurrence data poses it as an optimization problem in information theory — in this framework, an optimal clustering is one which minimizes the resulting loss in mutual information. A divisive clustering algorithm that monotonically reduces this loss function was recently proposed. In this paper we show that sparse high-dimensional data presents special challenges which can result in the algorithm getting stuck at poor local minima. We propose two solutions to this problem: (a) a "prior" to overcome infinite relative entropy values as in the supervised Naive Bayes algorithm, and (b) local search to escape local minima. Finally, we combine these solutions to get a robust algorithm that is computationally efficient. We present detailed experimental results to show that the proposed method is highly effective in clustering document collections and outperforms previous information-theoretic clustering approaches.

1 Introduction

Clustering is a central problem in unsupervised learning [8]. Presented with a set of data points, clustering algorithms group the data into clusters according to some notion of similarity between data points. However, the choice of similarity measure is a challenge and often an *ad hoc* measure is chosen. Information Theory comes to the rescue in the important situations where non-negative co-occurrence data is available. A novel formulation poses the clustering problem as one in information theory: find the clustering that minimizes the loss in (mutual) information [21, 6]. This information-theoretic formulation leads to a "natural" divisive clustering algorithm that uses relative entropy as the measure of similarity and monotonically reduces the loss in mutual information [6].

However, sparse and high-dimensional data presents special challenges and can lead to qualitatively poor local minima. In this paper, we demonstrate these failures and then propose two solutions to overcome these problems. First, we use a prior as in the supervised Naive Bayes algorithm to overcome infinite relative entropy values caused by sparsity. Second, we propose a local search strategy that is highly effective for high-dimensional data. We combine these solutions to get an effective, computationally efficient algorithm. A prime example of high-dimensional co-occurrence data is word-document data; we show that our algorithm returns clusterings that are better than those returned by previously proposed information-theoretic approaches.

The following is a brief outline of the paper. Section 2 discusses related work while Section 3 presents the information-theoretic framework and divisive clustering algorithm of [6]. The problems due to sparsity and high-dimensionality are illustrated in Section 4. We present our two-pronged solution to the problem in Section 5 after drawing an analogy to the supervised Naive Bayes algorithm in Section 5.1. Detailed experimental results are given in Section 6. Finally we present our conclusions and ideas for future work in Section 7.

A word about notation. Upper-case letters such as X, Y will denote random variables, while lower-case letters such as x, y denote individual set elements. $\hat{Y}$ denotes a random variable obtained from a clustering of Y while $\hat{y}$ denotes an individual cluster. Probability distributions will be denoted by p, q when the random variable is obvious or by $p(X)$, $p(X|y)$ to make the random variable explicit. Boldfaced letters, such as $\mathbf{y}, \hat{\mathbf{y}}$, will denote $p(X|y), p(X|\hat{y})$ for brevity. The logarithmic base 2 is used throughout this paper.

2 Related work

Clustering is a widely studied problem in unsupervised learning, and a good survey of existing methods can be found in [8, 13, 11]. Clustering algorithms can be categorized into agglomerative clustering algorithms and divisive clustering algorithms. Agglomerative clustering algorithm starts with each individual data item in its own cluster and iteratively merge clusters until all items belong in one cluster; while divisive clustering initially place all items in one cluster and clusters are repeatly split. When clusters are merged or split, some notion of similarity/distance measure is often defined. Similarity/distance measures between data items can be defined based on a measure, such as Euclidean distance, cosine [24] or based on boolean or categorical values [10]. Alternative clustering algorithms are based on graph partitioning [14].

For the case of non-negative, co-occurrence data, our information-theoretic framework is similar to the one used in the information bottleneck method [25]. The information bottleneck method tries to minimize the quantity $I(Y;\hat{Y})$ in order to gain compression in addition to maximizing the mutual information $I(X;\hat{Y})$; the optimization problem considered in [25] is $I(Y;\hat{Y}) - \beta I(X;\hat{Y})$ where β is the tradeoff between compression and preservation of mutual information. Information bottleneck algorithm yields a "soft" clustering of the data using a procedure which similar to the deterministic annealing [19] A greedy agglomerative hard clustering was used in [1, 23] to cluster words in order to reduce feature size for supervised text classification. For the same task, [6] proposed a divisive hard clustering algorithm that directly minimizes the loss in mutual information and was found to result in higher classification accuracies than [1, 23]. An agglomerative hard clustering version of information bottleneck algorithm is used in [22] to cluster documents after clustering words. The work in [9] extended the above work to repetitively cluster documents and then words. Methods based on sequential optimization are used in [2, 21]. As we demonstrate in Section 6, our proposed algorithm yields better clusterings than the above approaches, while being more computationally efficient.

Information-theoretic methods have been used for a variety of tasks in machine learning [19] including text classification [20]. Distributional clustering of words was first proposed in [18] and subsequently used by [1] for reducing the feature size for text classifiers. A general statistical framework for analyzing co-occurrence data based on probabilistic clustering by mixture models was given in [12].

3 Divisive Information-Theoretic Clustering

Let X and Y be two discrete random variables that take values in the sets $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$ respectively. Suppose that we know their joint probability distribution $p(X,Y)$; often this can be estimated using co-occurrence data. Consider the case where we want to cluster Y. Let $\hat{Y}$ denote the "clustered" random variable that ranges over the disjoint clusters $\hat{y}_1, \ldots, \hat{y}_k$, i.e.,

$$\cup_{i=1}^{k} \hat{y}_i = \{y_1, \ldots, y_n\}, \text{ and } \hat{y}_i \cap \hat{y}_j = \phi, \quad i \neq j.$$

Figure 1 illustrates the idea: by using $\hat{\mathbf{y}}_j$ as the representative for all y_i that are in cluster $\hat{y}_j$ we compress the large co-occurence matrix on the left side of the figure into the compressed matrix on the right. By this compression we will lose "information" but a good clustering will maintain as much "information" as possible in the compressed matrix. This subjective notion can be made objective by using the concept of mutual information from information theory, as we show below.