# *CoFD*: An Algorithm for Non-distance Based Clustering in High Dimensional Spaces*

Shenghuo Zhu, Tao Li, and Mitsuonri Ogihara

Department of Computer Science
University of Rochester
Rochester, NY 14620
{zsh,taoli, ogihara}@cs.rochester.edu

**Abstract.** The clustering problem, which aims at identifying the distribution of patterns and intrinsic correlations in large data sets by partitioning the data points into similarity clusters, has been widely studied. Traditional clustering algorithms use distance functions to measure similarity and are not suitable for high dimensional spaces. In this paper, we propose *CoFD* algorithm, which is a non-distance based clustering algorithm for high dimensional spaces. Based on the maximum likelihood principle, *CoFD* is to optimize parameters to maximize the likelihood between data points and the model generated by the parameters. Experimental results on both synthetic data sets and a real data set show the efficiency and effectiveness of *CoFD*.

## 1 Introduction

Clustering problems arise in many disciplines and have a wide range of applications. Intuitively, the clustering problem can be described as follows: let $W$ be a set of $n$ multi-dimensional data points, we want to find a partition of $W$ into clusters such that the points within each cluster are "similar" to each other. Various distance functions have been widely used to define the measure of similarity. The problem of clustering has been studied extensively in the database [20,13], statistics [5,7] and machine learning communities [8,12] with different approaches and different focuses.

Most clustering algorithms do not work efficiently in high dimensional spaces due to the *curse of dimensionality*. It has been shown that in a high dimensional space, the distance between every pair of points is almost the same for a wide variety of data distributions and distance functions [6]. Many feature selection techniques have been applied to reduce the dimensionality of the space [17]. However, as demonstrated in [2], the correlations in the dimensions are often specific to data locality; in other words, some data points are correlated with a given set of features and others are correlated with respect to different features.

* The project is supported in part by NIH Grants 5-P41-RR09283, RO1-AG18231, and P30-AG18254 and by NSF Grants EIA-0080124, NSF CCR-9701911, and DUE-9980943.

Y. Kambayashi, W. Winiwarter, M. Arikawa (Eds.): DaWaK 2002, LNCS 2454, pp. 52–62, 2002.

As pointed out in [15], all methods that overcome the dimensionality problems have an associated and often implicit or adaptive-metric for measuring neighborhoods.

In this paper, we present *CoFD*[1], a non-distance based algorithm for clustering in high dimensional spaces. The *CoFD* algorithm described here is an improvement over our previous method [22] and it contains several major extensions and revisions. The main idea of *CoFD* is as follows: Suppose that a data set $W$ with feature set $S$ needs to be clustered into $K$ classes, $C_1, \ldots, C_K$ with the possibility of recognizing some data points to be outliers. The clustering of the data then is represented by two functions, the *data map* $D : W \to \{0, 1, \ldots, K\}$ and the *feature map* $F : S \to \{0, 1, \ldots, K\}$, where $1, \ldots, k$ correspond to the clusters and 0 corresponds to the set of outliers. Accuracy of such representation is measured using the log likelihood. Then, by the *Maximum Likelihood Principle*, the best clustering will be the representation that maximizes the likelihood. In *CoFD*, several approximation methods are used to optimize $D$ and $F$ iteratively. The *CoFD* algorithm can also be easily adapted to estimate the number of classes when the value $K$ is not given as a part of the input. An added bonus of *CoFD* is that it produces interpretable descriptions of the resulting classes since it produces an explicit feature map. The rest of the paper is organized as follows: section 2 introduces the core idea and presents the details of *CoFD*; section 3 shows our experimental results on both the synthetic data sets and a real data set; section 4 surveys the related work; finally our conclusions and directions for future research are presented in section 5.

## 2 The *CoFD* Algorithm

This section describes *CoFD* and the core idea behind it. We first present the *CoFD* algorithm for binary data sets. Then we will show how to extend it to continuous or non-binary categorical data sets in Section 2.4.

### 2.1 The Model of *CoFD*

Suppose we wish to divide $W$ into $K$ classes with the possibility of declaring some data as outliers. Such clustering can be represented by a pair of functions, $(F, D)$, where $F : S \to \{0, 1, \ldots, K\}$ is the *feature map* and $D : W \to \{0, 1, \ldots, K\}$ is the *data map*.

Given a representation $(F, D)$ we wish to be able to evaluate how good the representation is. To accomplish this we use the concept of *positive features*. Intuitively, a positive feature is one that best describes the class it is associated with. Suppose that we are dealing with a data set of animals in a zoo, where the vast majority of the animals is the monkey and and the vast majority of the animals is the four-legged animal. Then, given an unidentified animal having four legs in the zoo, it is quite natural for one to guess that the animal is a monkey

[1] *CoFD* is the abbreviation of Co-training between Feature maps and Data maps.