Design and Analysis of Subspace Clustering Algorithms and their Applicability.

Jyoti D. Pawar,

Advisor : Dr. P. R. Rao,

Goa University, INDIA.

INTRODUCTION

What is Subspace Clustering?

The problem of **automatically identifying clusters present in the subspaces of a high dimensional data** space that allows better clustering of the data points than the original space.

Why Subspace Clustering?

♦ Most of the clustering algorithms have been designed to discover clusters in the full dimensional space. Hence they are **not effective in identifying clusters that exist in the subspaces** of the original data space.

♦ Many times the **data records contain some missing values**. Such missing values are normally replaced with values taken from a distribution.

♦ The clustering results produced by most of the clustering algorithms depend a lot on **the order in which input records are processed.**

Applications -

♦ **Sales analysis** - by identifying the different subspace clusters that exist in the huge amount of sales data, we can find which of the different attributes are related. This can be useful in promoting the sales and in planning the inventory levels of different products.

♦ It can be used for finding the subspace clusters **in spatial databases** and some useful decisions can be taken based on the subspace clusters identified.

♦ It can also be used for **indexing OLAP data.**

♦ and more **???** .(*this needs to be figured out after studying some more real life applications and from the feedback obtained*)

A TYPICAL SUBSPACE CLUSTERING ALGORITHM – *HOW DOES IT WORK?*

An overview –

The algorithm consists of **three main steps** namely

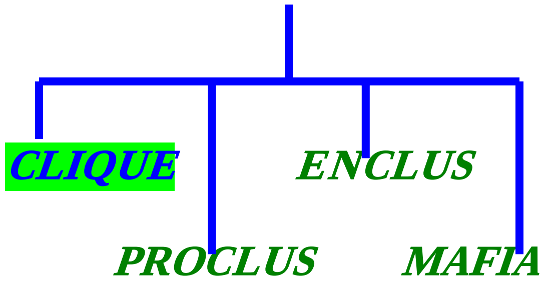
1. ***Identification of the subspaces that contain clusters.***
2. ***Identification of clusters.***
3. ***Generation of minimal description for the clusters.***

Step 1 involves **finding the dense units in the different subspaces** and is the most time consuming one. The **time complexity** of this step is **$O(c^k+mk)$** for a constant c, where k is the highest dimensionality of any dense unit and m the number of the input points. The algorithm is based on the level-wise **Apriori algorithm** and makes *k passes* over the database.

Step 2 involves using the depth-first search algorithm for **finding the connected components** in a graph using the dense units as vertices, and having an edge iff two dense units share a common face.

Step 3 generates a **concise description of the cluster** with the help of the connected components identified in *step 2.*

Examples of some subspace clustering algorithms



OBJECTIVES AND METHODOLOGY USED

a) To improve the efficiency of the step 1 of the subspace clustering algorithm

Possible directions :

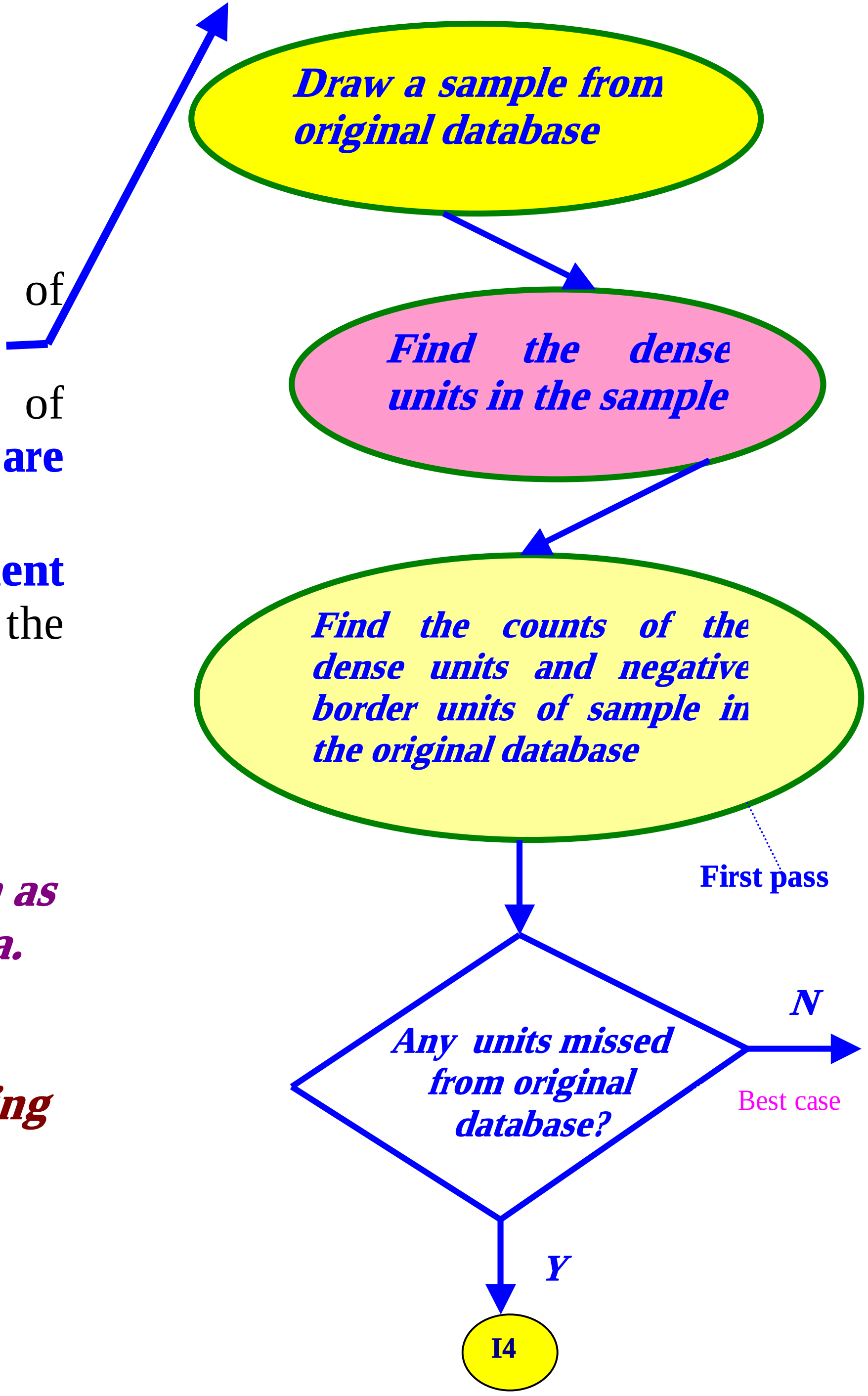
- **reduce** the number of **passes over the data**.
- **reduce** the number of **candidates which are counted**.
- **devise** a more **efficient method** to find the dense units.

b) To design a subspace clustering algorithm for use in applications such as analysis of census data.

c) To develop a subspace clustering algorithm using support constraints.

d) To design a subspace clustering algorithm for streaming data.

Used sampling -



Draw a sample from original database

Find the dense units in the sample

Find the counts of the dense units and negative border units of sample in the original database

First pass

Any units missed from original database?

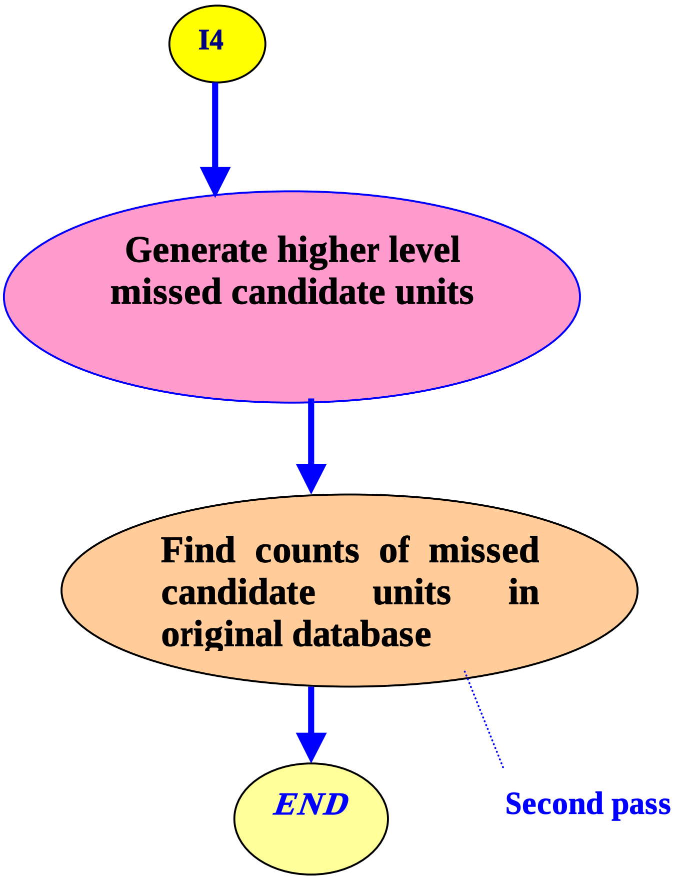
N

Best case

Y

I4

METHODOLOGY USED (CONTD..)



I4

Generate higher level missed candidate units

Find counts of missed candidate units in original database

END

Second pass

OBSERVATIONS.

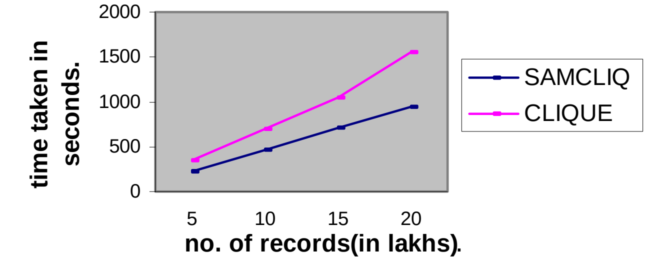


Figure 1: Scalability with the number of records.

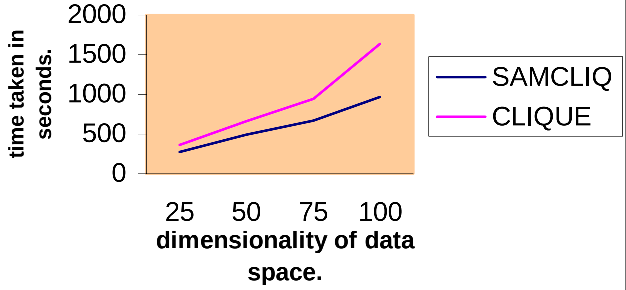


Figure 2 : Scalability with the dimensionality of data space

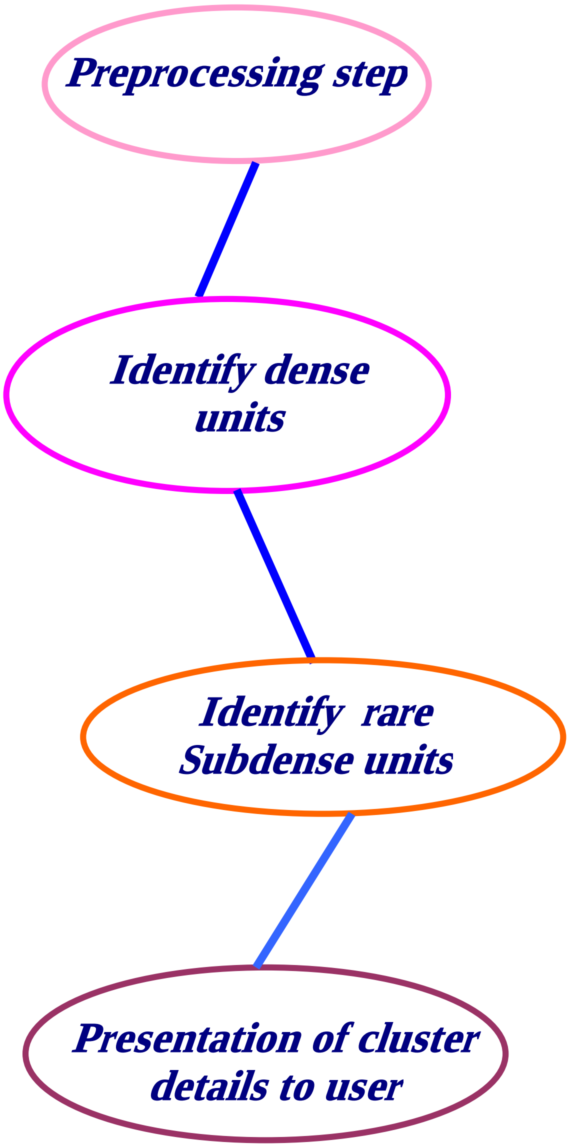
METHODOLOGY USED (CONTD…)

To design a subspace clustering algorithm for use in applications such as analysis of census data.

METHOD USED-

Properties and requirements –

- **data of mixed types.**
- **data attributes occur with different levels of frequency**
- **not very high dimensional clusters found.**
- **need to detect occurrences of rare/infrequent attribute values in the subspace clusters of original data.**
- **need for a better way to present the cluster details.**



Preprocessing step
Identify dense units
Identify rare Subdense units
Presentation of cluster details to user