Распознавание цифр

Задача
Определить число кластеров и разбить цифры на кластеры.
Предложить интерпретацию кластеров.

Каждая строка набора данных описывает цифру.
Цифры отсканированы с ошибками
В обучающей выборке присутствует группирующая переменная - правильная цифра.

Окончательной целью задачи является модель для распознавания цифр.
И это не задача кластеризации.

Но применим такой вариант решения.
Первый этап. Сначала кластеризуем наблюдения, чтобы похожие цифры собрались в группы.
Второй этап. Для каждого кластера строим свою модель. Придется строить больше моделей, зато каждая из них будет решать более простую задачу, строить их будет легче, общее качество распознавания увеличится.

В лабораторной работе Вам надо решить задачу первого этапа.
Определить число кластеров и разбить цифры на кластеры. Также надо предложить интерпретацию для каждого кластера.

Группирующую переменную "A" нельзя использовать при кластеризации,
но рекомендуется использовать ее при интерпретации кластеров.

В данных 7 переменных с именами "B" - "H",
измеренных в номинальной шкале
0 = линия присутствует
1 = линия отсутствует

Линии соответствуют черточкам на экране калькулятора

B - top horizontal,
C - upper left vertical,
D - upper right vertical,
E - middle horizontal,
F - lower left vertical,
G - lower right vertical,
H - bottom horizontal.

В наборе данных 8 переменных и 500 наблюдений
По неизвестной причине в таблице данных каждый столбец присутствует дважды

Повторим ниже первые 10 строчек набора данных

A	B	C	D	E	F	G	H
7	1	0	1	0	0	1	0
1	0	0	1	0	0	1	0
4	0	1	1	1	0	1	0
2	1	1	1	1	1	0	0
8	0	1	1	1	1	1	1
1	0	0	1	0	0	1	0
5	1	1	0	1	0	1	1
6	1	0	0	1	1	1	1
2	1	0	1	1	1	0	1
8	1	1	1	1	0	1	1

Данные заимствованы из книги
Breiman, Friedman, Olshen, Stone (1984). Classification and regression trees.