Восстановление и кластеризация данных.

Порядок работы:

1) Сформировать датасеты с помощью программы из 1 лабораторной работы разных величин (малый ~ 30000, средний ~ 75000, большой ~ 250000)
2) Вычислить среднее значение, значение медианы и моды, построить распределение (во всех датасетах).
3) Удалить в столбцах 3, 5, 10, 20, 30 % значений, притом делать это выбросами (То есть удаляем кусками/матрицами 2x2, 4x2, 3x3 и тд).
4) Выполнить заполнение пропусков методами.
5) Произвести кластеризацию и классификацию полученного датасета по выбранному варианту.
6) Выбрать наиболее информативные признаки для датасетов и повторить кластеризацию.
7) Оценить результаты: вычислить среднее значение, значение медианы и моды, построить распределение и сравнить их, а также сравнить кластеризацию и классификацию восстановленного и оригинального датасета.

A1.Методы заполнения пропусков

1) Удаление строк с пустыми значениями - Анализ полных наблюдений
2) Попарное удаление
3) Хот-Дек
4) Метод подстановки с подбором внутри групп.
5) Метод заполнения средним значением
6) Метод заполнения значением медианы
7) Метод заполнения моды
8) Повторение результата последнего наблюдения перед пропуском
9) Заполнение пропусков на основе линейной регрессии
10) Заполнение пропусков на основе стохастической линейной регрессии
11) Метод восстановления пропущенного значения сплайн-интерполяцией по присутствующим элементам
12) Метод восстановления пропущенного значения на основе Zet-алгоритма

Примечание **- выбора наиболее эффективного алгоритма**

Эффективность того или иного метода устанавливается экспериментально в такой последовательности:

1) формируется массив комплектных записей. Для этого неполные наблюдения исключаются из рассмотрения;
2) «искусственно» создаются пропущенные значения, т. е. в таблице удаляются некоторые элементы $a_1, ..., a_n$;
3) пропущенные значения поочередно предсказываются с использованием разных методов $M_1, ..., M_k$;
4) рассчитываются суммарные относительные погрешности для каждого метода:

$$\Delta_{M_j} = \sum_{i=1}^{n} \frac{|a_i - \overline{a}_i|}{a_i} \cdot 100\ \%,$$

где j — номер метода, j =1, ..., k ; a_i — пропущенное значение; $\overline{a_t}$ - предсказанное значение.

Метод, для которого суммарная относительная погрешность будет минимальной, будет наиболее эффективным.

Поскольку набор методов заполнения пропусков достаточно велик, то тестировать можно те методы, которые реализованы в пакете обработки данных, используемом исследователем, поскольку проверка эффективности алгоритмов данных без применения программного обеспечения — трудоемкая задача.

Сравнение эффективности алгоритмов

Процесс сравнения эффективности алгоритмов, можно организовать следующим образом:

1) сформировать датасет, состоящий только из полных наблюдений;
2) оценить на нем параметры распределений переменных и осуществить интересующие виды анализа. Полученные только на полных наблюдениях результаты станут эталоном − базой для дальнейших сравнений;
3) внести в датасет полных наблюдений различное количество случайных пропусков. В итоге получится совокупность массивов данных разной степени полноты;
4) на каждом из них повторить шаг 2 и оценить смещения в оценках, возникшие в результате наличия в данных разного количества пропусков;
5) заполнить в каждом датасете пропуски с использованием нескольких сравниваемых методов заполнения пропусков. Повторить на каждом полном массиве шаг 2.
6) сравнить полученные при использовании каждого из методов заполнения пропусков результаты с эталонными результатами, полученными на шаге 2.
7) выбрать для каждой ситуации наиболее эффективный метод заполнения пропусков и отразить их в отчете.

А2.Методы кластеризации

1) Иерархическая
2) Алгоритм максиминного расстояния
3) Алгоритм ISODATA
4) Алгоритмы класса ISODATA
5) Алгоритм CURE (Clustering Using REpresentatives)
6) Алгоритм ФорЭл (FOREL)

А3.Методы выбора наиболее информативные признаков

1) Компактность (плотность)
2) Разнесенность образов в пространстве характеристик
3) Метод последовательного сокращения (алгоритм Del)
4) Метод последовательного добавления признаков (алгоритм Add)
5) Метод случайного поиска с адаптацией (алгоритм СПА)

А4.Методы способа измерения расстояния

1) Евклидово расстояние
2) Квадрат евклидового расстояния
3) Корреляция Пирсона
4) Расстояние Чебышева (Chebychev)
5) Расстояние Минковского (Minkowski)

А5.Вычисления расстояний между кластерами

1) Ближайший сосед (Nearest neighbor);
2) Самый дальний сосед (Furthest neighbor);
3) Медианная кластеризация (Median clustering);
4) Метод Варда (Ward-Method):

Примечание - Будем использовать расстояние Уорда для определения расстояния между множествами точек W и S

$$R_{ward}(W,S) = \frac{|W|\cdot|S|}{|W|+|S|}\cdot\rho\left(\sum_{w\in W}\frac{w}{|W|},\sum_{s\in S}\frac{s}{|S|}\right)$$

И формулу Ланса-Уильямса для определения расстояния между объединением множеств точек W = U ∪ V и множеством точек S

$$R_{lnwl}(U\cup V,S) = \alpha_U\cdot R_{ward}(U,S) + \alpha_V\cdot R_{ward}(V,S) + \beta\cdot R_{ward}(U,V)$$

Где

$$\alpha_U = \frac{|U|+|S|}{|W|+|S|}\ ;\ \alpha_V = \frac{|V|+|S|}{|W|+|S|}\ ;\ \beta = \frac{-|S|}{|W|+|S|}$$

А6.Методы оценки качества кластеризации

1) Индекс Rand
2) Индекс Жаккара
3) Индекс Фоулкса – Мэллова
4) Индекс Phi
5) Компактность кластеров
6) Отделимость кластеров