1　実験目的・課題

- 単純な二次元データの分類
- 動物園の動物データ (多次元データ) の分類
- 食べ物の画像データの分類

を行う。

2　実装方法

2.1　クラスタリングの手法について

データのクラスタリングをする手法には以下のようなものがある。

- 最短距離法
- 最長距離法
- 重心法
- 群平均法
- K-means 法

最短距離法とは 2 つのクラスター同士について、最も距離が小さいものから順にグループにしていくものである。計算量が少ないが、ある 1 つのクラスタにひとつづつ吸収されてしまう鎖効果が起こってしまうことがある。

最長距離法はクラスター間の距離をそのクラスター間の中で最も遠いデータとし、その距離の小さいものからクラスターにしていく手法である。分類感度は高いがクラスター同士が離れてしまう拡散現象が起きることがある。

重心法はクラスターに属するデータの重心を決定し、重心間の距離の小さい順にクラスターを形成していく手法である。計算量は多いが分類感度が良い。

群平均法は 2 つのクラスターに含まれるすべてのデータ間の距離を計算し、その平均の値をクラスター間の距離としてクラスタリングをする手法である。鎖効果や拡散現象をおこしにくい。

K-means 法とはデータ数を n、クラスタ数を k として、以下のような手順で表されるアルゴリズムである。

1. 各データ $d_i(i = 1, \ldots, n)$ に対してランダムにクラスタを割り振る
2. 各クラスタの中心 $m_j(j = 1, \ldots, k)$ を求める
3. 各 d_i と各 m_i との距離を求め d_i を最も近いクラスタに割り当てなおす
4. クラスタの割り当てが変化しなかった場合、終了する。
   そうでない場合手順 2 に戻る

結果は、最初に割り振られたクラスタに依存し、1 回の結果で最良のものが得られるとは限らない。

2.2 二次元データの分類

2.2.1 最短距離法

二次元データ間の距離をユークリッド距離 $d = \sqrt{x^2 + y^2}$ で定義する。これを最短距離法で分類することを考える。最短距離法では単に短い距離のものから順にクラスタにまとめていけばよい。これはクラスカル法で最小全域木を構築するのと同様の手順で行うことができる。

クラスカル法とはグラフ理論において重み付き連結グラフの最小全域木を求めることが出来る以下のようなアルゴリズムのことである。

1. $cost(e_0) \leq cost(e_1) \leq \ldots \leq cost(e_m)$ が成り立つように辺を並び替える
2. $T := (V(G), \emptyset)$ とする
3. i=1..m について、$E(T) \cup \{e_i\}$ が閉路を含まないならば $E(T) := E(T) \cup \{e_i\}$ とする

二次元データを頂点、各点間の距離を重みとした辺からなるグラフを考える。最初は N 個の木 (森) が存在し、このアルゴリズムが終了したときは 1 つの木だけが残っている。この手順の途中で木の数が k になったときにアルゴリズムを終了すると、k 個のクラスタに分類することが出来ている。クラスタの表現には Union Find というデータ構造を使用する。これを実装したものをソースコード 1 に示す。

ソースコード 1　Rust による最短距離法の実装

```
fn min_dist_method(data: &Vec<Point>, m: usize) -> Vec<usize> {
    let n = data.len();
    let mut edges: Vec<(f64, usize, usize)> = Vec::new();
    for i in 0..n {
        for j in i + 1..n {
            edges.push((calc_dist(data[i], data[j]), i, j));
        }
    }
    edges.sort_by(|a, b| a.partial_cmp(b).unwrap());
    let mut uni = Unionfind::new(n);
    for (_, i, j) in edges {
        if uni.tree_count == m {
            break;
        }
        uni.unite(i, j);
    }
    let mut cluster = (0..n).into_iter().map(|i| uni.root(i)).collect::<Vec<usize>>();
    cluster = compression(cluster);
    cluster
}
```

2.2.2 群平均法

群平均法ではクラスタ c_a と c_b の距離を c_a の大きさ (所属しているデータの数) を N、c_b の大きさを M として、$\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} d(c_{a_i}, c_{b_j})$ として計算する。ここで、$d(p_i, p_j)$ はデータ間の距離を計算する関数であ

る。こうして計算された距離の中で最小のものを併合していく。

クラスタはデータの追加が出来る集合とそれを管理する削除が出来る集合が必要であるため、HashMap で Vec を管理する。

ソースコード 2　群平均法の実装

```
fn group_average_method(data: Vec<Point>, m: usize) -> HashMap<usize, Vec<Point>> {
    let mut cluster: HashMap<usize, Vec<Point>> = HashMap::new();
    let n = data.len();
    for i in 0..n {
        let vi = vec![data[i]];
        cluster.insert(i, vi);
    }
    let mut indexes = (0..n).into_iter().collect::<Vec<_>>();
    while cluster.len() > m {
        let (mut mi, mut mj) = (0, 0);
        let mut mdist = 1e10;
        for i in &indexes {
            for j in &indexes {
                if i < j {
                    let dist = calc_dist_average(&cluster[&i], &cluster[&j]);
                    if mdist > dist {
                        mdist = dist;
                        mi = *i;
                        mj = *j;
                    }
                }
            }
        }
        let mut clst_j = cluster.remove(&mj).unwrap();
        cluster.get_mut(&mi).unwrap().append(&mut clst_j);
        let rmidx = indexes.iter().position(|x| *x == mj).unwrap();
        indexes.remove(rmidx);
    }
    cluster
}
fn calc_dist_average(a: &Vec<Point>, b: &Vec<Point>) -> f64 {
    let mut sum_d: f64 = 0.;
    for ai in a {
        for bj in b {
            sum_d += calc_dist_point(*ai, *bj);
        }
    }
    sum_d / ((a.len() * b.len()) as f64)
}
```

2.2.3 k-means 法

k-means 法は初期状態によって結果が偏ってしまうことがあるので複数回実行して分散の値が小さいものを結果が良かったものとして採択する。分散の値は、各クラスタについて重心とのユークリッド距離で計算する。重心の座標は次元ごとに平均をとったものとする。その座標との距離の平均を d_{mean} とし、各クラスタについて $\sum_{i=1}^{N}(d(center, p_i) - d_{mean})^2$ を計算し、その値の平均値を分散とした。

ソースコード 3　k-means 法の実装

```
fn k_means(data: &Vec<Point>, m: usize) -> (Vec<Vec<Point>>, Vec<usize>) {
    let n = data.len();
    let mut rng = rand::thread_rng();
    let mut cluster = (0..n).into_iter().map(|_| rng.gen_range(0..m)).collect::<Vec<_
        >>();
    let mut center = calc_center(&data, &cluster, m);
    loop {
        let mut new_cluser = vec![0; n];
        for i in 0..n {
            let mut d_min = 1e10;
            for j in 0..m {
                let d_cur = calc_dist(data[i], center[j]);
                new_cluser[i] = if d_min > d_cur {
                    d_min = d_cur;
                    j
                } else {
                    new_cluser[i]
                };
            }
        }
        if cluster == new_cluser {
            break;
        }
        cluster = new_cluser;
        center = calc_center(&data, &cluster, m);
    }
    let mut cluster_set: Vec<Vec<Point>> = vec![Vec::new(); m];
    for i in 0..n {
        cluster_set[cluster[i]].push(data[i]);
    }
    (cluster_set, cluster)
}
```

2.3 多次元データの分類

二次元データの分類の時と同様にクラスカル法を実行することで最短距離法での分類を行う。このとき n 次元データ a と b の間の距離は $d_{ab} = \sqrt{\sum_{i=1}^{n} w_i (a_i - b_i)^2}$ として計算した。ここで w_i は各系列に対する重みである。ここで重みはランダムに決定し何度か実行し、正解との差が小さいものを探した。

2.4 画像データの分類

画像データを次元の小さい多次元データに圧縮することで、前節と同様に分類をすることができるようになる。

画像のデータは Python の OpenCV では RGB を表す 3 要素からなる配列の二次元配列である。この RGB の値は $256^3 = 16777216$ 通り存在する。各色についてその色が画像にいくつ含まれているかをカウントした 16777216 次元データを比べることで画像間の距離とすることでクラスタリングに必要な情報が得られる。しかし、16777216 個のデータを比較するのはデータの量が多すぎるため困難である。ここで、各画素の値を 64 で割り、0 から 255 の値を 0 から 3 までの値に変換する。こうすることで各画素を $4^3 = 64$ 通りの色に圧縮することが出来る。

画像を 64 次元データに圧縮することができたので最短距離法によって画像の分類を実行する。

画像の分類結果を確認するために、Python の subprocess パッケージの run 関数を使い、分類のたびに新たにディレクトリを作成し、そこへ元画像のファイルをコピーした。

3 結果と考察

3.1 二次元データの分類

最短距離法、群平均法、k-means 法のそれぞれでクラスタ 4 つ、分散 200 のデータセットをクラスタリングした結果を図 1、2、3 に示す。

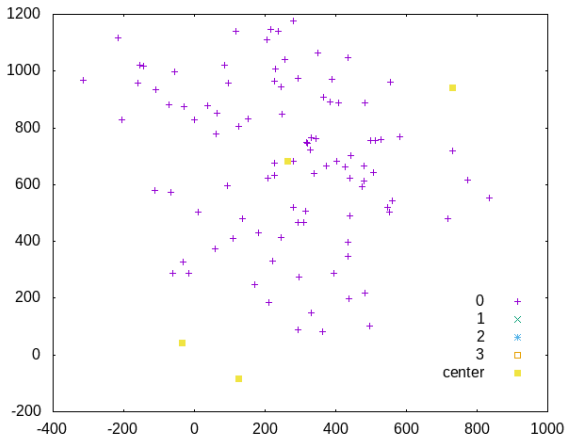

図 1　最短距離法によるクラスタリング

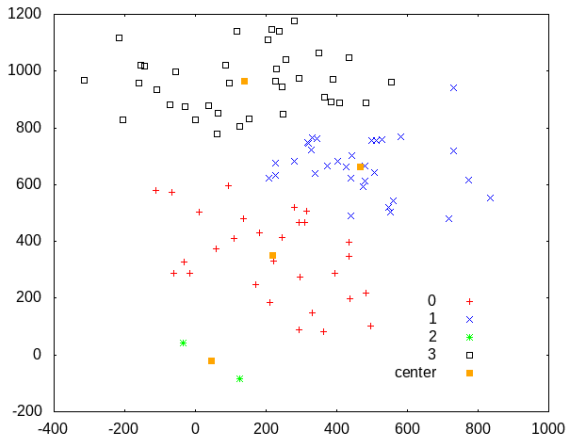

図 2　群平均法によるクラスタリング

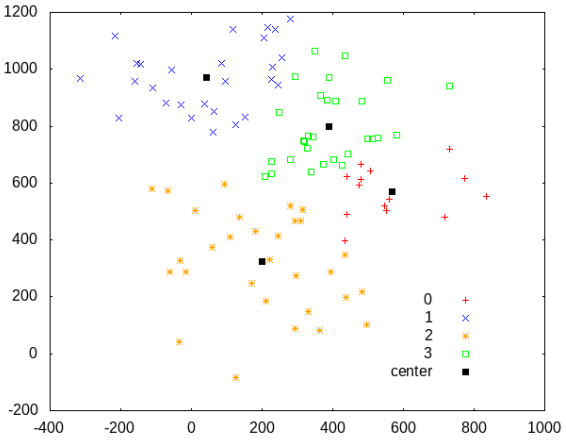

図 3　k-means 法によるクラスタリング

最短距離法によるクラスタリングでは、クラスタ 1,2,3 が 1 つの点からなるクラスタになってしまっている。これは、鎖効果が起こってしまったためだと考えられる。群平均法ではこれが改善され、k-means 法ではさらに改善されて均等な 4 つのクラスタに分類できていることがわかる。

3.2　多次元データの分類

3.3　画像データの分類

参考文献

[1] クラスター分析の手法（階層クラスター分析）
https://www.albert2005.co.jp/knowledge/data_mining/cluster/hierarchical_clustering
[2] k-means 法を理解する
https://qiita.com/g-k/items/0d5d22a12a4507ecbf11