1.1. Матрице сличности (Proximity matrices)

Подаци су понекад представљени директно у смислу близине (сличности или афинитета) између парова објеката. То могу бити било њихове сличности или разлике (разлика или недостатак афинитета). На пример, у експериментима из друштвених наука, од учесника се тражи да просуде по томе колико се одређени објекти међусобно разликују. Тада се разлике могу израчунати рачунањем просека тако прикупљених пресуда.

Узмите било коју слику и покушајте да обележите регионе које видите. Ако покуша још неко да уради исто, видећете да постоји извесна разлика у обележеним регионима. Шта је онда *истина* (ground-truth)?

Овај тип података може да се представи $N \times N$ матицом D, где је N број објеката, а сваки елемент $d_{ii'}$ представља близину између објеката i и i'. Ова матрица представља улаз у алгоритам за кластеризацију.

Већина алгоритама претпоставља да се на улазу добија **матрица различитости** са ненегативним елементима и нула елементима на дијагонали: $d_{ii} = 0, i = 1,2, \dots, N$. Ако су изворни подаци прикупљени као сличности, може се користити одговарајућа монотоно опадајућа функција за њихово претварање у различитост (погледајте пример за Фејсбук где је узета функција $f(x) = \frac{1}{x}$). Такође, већина алгоритама подразумева коришћење симетричне матрице различитости, па ако оригинална матрица D није симетрична, тада се она мора заменити са $(D + D^T)/2$.

Субјективно процењене различитости ретко су када **растојање** у строгом смислу (дистанца), јер неједнакост троугла $d_{ii'} \leq d_{ik} + d_{i'k}, \forall k \in \{1,2, \dots, N\}$ не важи. Стога се неки алгоритми који претпостављају **растојања** не могу користити са таквим подацима.

1.2. Различитост заснована на атрибутима

Доста често као податке имамо обсервације x_{ij} за $i = 1,2, \dots, N$ над променљивама $j = 1,2, \dots, p$ (*атрибути*). Имајући у виду да већина алгоритама за кластеризацију података узима матрицу различитости као улаз, на почетку прво морамо да конструишемо различитости између свих обсервација. У већини случајева најпре се дефинише различитост $d_j(x_{ij}, x_{i'j})$ између вредности j-ог атрибута, и тада се дефинише:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$$

као различитост између објеката i и i'. Најчешћи избор који се користи у литератури је квадрат дистанце:

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

Такође, и други избори су могући и они потенцијално могу да воде ка различитим резултатима. За неквантитативне атрибуте (нпр. категоријске променљиве), квадрат растојања није прави избор. Додатно, некада је пожељно додати тежине за различите атрибуте пре него да се свима додели иста важност.

Квантитативне променљиве

Мерења овог типа променљивих или атрибута су представљени континуалним реалним бројевима. Природно је да се разлика између њих дефинише као монотоно растућа функција њихове апсолутне разлике:

$$d(x_i, x_{i'}) = l(|x_i - x_{i'}|)$$

Поред квадратне грешке $(x_i - x_{i'})^2$, доста често се користи и само апсолутна грешка. За квадратну грешку већи је акценат на велике разлике, него на оне мале. Алтернативно, кластеризација се може заснивати и на корелацији:

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}},$$

где је $\bar{x}_i = \sum_j x_{ij}/p$. Битно је напоменути да се просек тражи на нивоу променљивих, а не на основу свих обсервација.

Ако су обсервације стандардизоване, тада важи:

$$\sum_j (x_{ij} - x_{i'j})^2 \; \alpha \; 2\big(1 - \rho(x_i, x_{i'})\big).$$

Због тога је кластеризација заснована на корелацији (сличност) еквивалентна оној кластеризацији заснованој на квадратној удаљености (различитост).

Ординалне променљиве

Вредности ове врсте променљивих су често представљене као суседне целобројне вредности, а посматране вредности се сматрају уређеним скупом. Примери су академске оцене (А, Б, Ц, Д, Ф), степен преференција (не може да поднесе, не воли, у реду, свиђа, сјајно).

Подаци о рангу су посебна врста уређених података. Мере грешака за ординалне променљиве су генерално дефинисане заменом њихових M оригиналних вредности са:

$$\frac{i-1/2}{M}, i = 1, \dots, M$$

по прописаном редоследу њихових изворних вредности. Затим се на овој скали третирају као квантитативне променљиве!

Категоријске променљиве

Са неуређеним категоријским (номиналним) променљивама, степен разлике између парова вредности мора бити експлицитно назначен. Ако променљива има M различитих вредности, оне могу бити организоване у $M \times M$ симетричној матрици са елементима $L_{rr'} = L_{r'r}, L_{rr} = 0, L_{rr'} \geq 0$. Најчешћи избор је $L_{rr'} = 1$ за све $r \neq r'$.

Различитост објеката када су подаци различитог типа

Даље дефинишемо поступак за комбиновање различитости p-индивидуалних атрибута $d_j(x_{ij}, x_{i'j}), j = 1,2, \dots, p$ у једну обједињену меру различитости $D(x_i, x_{i'})$ између објеката/обсервација $(x_i, x_{i'})$ који поседују одговарајуће вредности атрибута.

То се готово увек ради помоћу пондерисаног просека (конвексна комбинација).

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} w_j \cdot d_j(x_{ij}, x_{i'j}); \ \sum_{j=1}^{p} w_j = 1\,.$$

Овде је w_j пондер додељен j-том атрибуту који регулише релативни утицај те променљиве у одређивању укупне разлике међу објектима. Овај избор треба да се заснива на разматрању предмета.

Ако је циљ откривање природних група у подацима, неки атрибути могу показивати већу тенденцију груписања од других. Променљивама које су релевантније за раздвајање група треба доделити већи утицај у дефинисању различитости објеката. Давање свим атрибутима подједнаки утицај, у овом случају настојаће да заклони групе које се траже до те мере да их алгоритам за кластеризацију података не може открити. Слика 14.5 приказује пример (из књиге *The elements of statistical learning*).

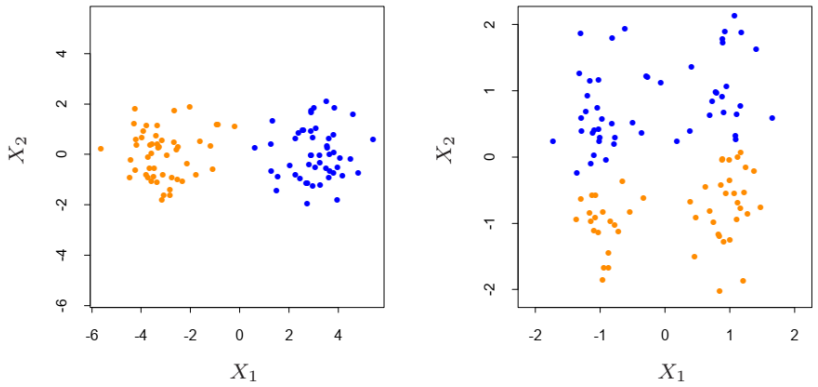

FIGURE 14.5. *Simulated data: on the left, K-means clustering (with K=2) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights* $1/[2 \cdot \mathrm{var}(X_j)]$. *The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.*

Иако једноставни генерички рецепти за одабир појединачних атрибутних разлика $d_j(x_{ij}, x_{i'j})$ и њихове тежине w_j могу дати солидне резултате, не постоји замена за пажљиво размишљање у контексту сваког појединачног проблема. **Одређивање одговарајуће мере несличности је далеко више важан корак за постизање успеха у кластеризацији од избора алгоритма за кластеризацију**. Овај аспект проблема је наглашен мање у литератури о кластеризацији него самим алгоритмима, јер зависи од специфичности доменског знања и мање је подложан општем истраживању.

Додатни ресурси:

- https://datascience.stackexchange.com/questions/22/k-means-clustering-for-mixed-numeric-and-categorical-data

„You should not use k-means clustering on a dataset containing mixed datatypes. Rather, there are a number of clustering algorithms that can appropriately handle mixed datatypes. Some possibilities include the following:

1) Partitioning-based algorithms: k-Prototypes, Squeezer

2) Hierarchical algorithms: ROCK, Agglomerative single, average, and complete linkage

3) Density-based algorithms: HIERDENC, MULIC, CLIQUE

4) Model-based algorithms: SVM clustering, Self-organizing maps

If you would like to learn more about these algorithms, the manuscript 'Survey of Clustering Algorithms' written by Rui Xu offers a comprehensive introduction to cluster analysis."

- https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3

„Most similar and dissimilar clients according to Gower distance"

- https://towardsdatascience.com/clustering-datasets-having-both-numerical-and-categorical-variables-ed91cdca0677

„This article discusses a clustering approach using Gower distance, the PAM (Partitioning Around Medoids) method, and silhouette width and explains each of the steps with an implementation in R."

1.3. K-means имплементација: Пример

Додатни примери:

- Demonstration of k-means assumptions: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html#sphx-glr-auto-examples-cluster-plot-kmeans-assumptions-py
- Selecting the number of clusters with silhouette analysis on k-means clustering: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

1.4. K-medoids algorithm

PowerPoint prezentacija

1.5. Expectation-Maximization (EM)

EM алгоритам (пдф)

1.6. Нови алгоритам за кластеризацију података: Affinity propagation

- https://warwick.ac.uk/fac/sci/dcs/research/combi/seminars/freydueck_affinitypropagation_science2007.pdf
- https://en.wikipedia.org/wiki/Affinity_propagation

ОБЈАШЊЕЊЕ корака алгоритма кроз пример:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.490.7628&rep=rep1&type=pdf

БЛОГ:

https://towardsdatascience.com/unsupervised-machine-learning-affinity-propagation-algorithm-explained-d1fef85f22c8