3.3.1 Методи за квантово клъстериране

Клъстерирането като техника за машинно обучение може да се използва освен за класификация и разпознаване на шаблони, и за допълване на шаблони, както и за асоциативна памет. Поради природата на квантовия компютър използването на клъстериране за допълване на шаблони и за асоциативна памет е силно недетерминистично в контекста на изследванията в дисертационния труд и затова квантовото клъстериране се разглежда основно като техника за класификация и разпознаване на шаблони.

За разлика от разгледаното в т. **3.2** преобразуване на алгоритъма k-means в квантов еквивалент, който работи с двоични вектори, при разглежданата тук задача се работи със стандартни входни данни, т.е. не с дескриптори на изображение, а потребителски имена и пароли. За k-means отново се използва (12) и се конструира следното състояние:

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|\vec{x}\rangle|0\rangle + \frac{1}{\sqrt{N}}\sum_{j=1}^{N}|\overrightarrow{y_j^S}\rangle|j\rangle), \tag{32}$$

където S текущият клъстер за множество от N референтни вектора $\{|y_j^S\rangle\}$ с дължина P и входен вектор $\overrightarrow{|x\rangle}$. Това е именно стъпката за назначаване. Разстоянието се изчислява ефективно с грешка $\epsilon = O(\epsilon^{-1}\log NP)$. Ако се използва отново swap-тестът за измерване на дистанцията, като сега се запазва метриката чрез Евклидово разстояние, то ще се получат следните състояния, базирани на (28):

$$|\phi\rangle = \frac{1}{\sqrt{Z}}(|x||0\rangle - \frac{1}{\sqrt{N}}\sum_{j=1}^{N}|y_j^S||j\rangle), \tag{33}$$

където $Z = |x|^2 + (\frac{1}{N})\sum_j |y_j|^2$.

Това отново се повтаря за всеки клъстер до конвергенция, достигане на достатъчна увереност. Всъщност, както стана ясно от предходната методология, може да бъде или система за препоръка, или оператор - човек. Самият алгоритъм k-means спада към типа клъстериращи алгоритми, които изграждат структурата си чрез порциониране или разделяне на входните данни. Друга, съвършено различна по природа технология за клъстериране, е агломеративното, което е разновидност на йерархичните алгоритми за клъстериране. Подробното му квантово представяне е разгледано в [Kong et al, 2017]. За конкретната задача ще бъде представено самото трансформиране на агломеративното клъстериране като алтернатива на разгледания в предишната точка подход за квантово клъстериране с алгоритъм k-means. Агломеративното клъстериране е възходящ подход[1], където всеки входящ обект е свой собствен първоначален клъстер. Преминаването нагоре в йерархията води до сливане на двойки клъстери, докато всички обекти не бъдат назначени в един такъв. Агломеративното клъстериране може да се съпостави с k-means алгоритъма по следния начин: приемат се N на брой входни

[1] Бел. пр. от англ. Bottom-up approach

елементи, а броят терминиращи клъстери също може да бъде зададен като k – в случая това не е задължителен белег, тъй като агломеративното клъстериране на изисква предварително зададен брой центроиди, самият алгоритъм определя колко клъстера да има в даден момент от времето. При условие, че всеки входящ обект принадлежи на своя клъстер, се изисква изчисляване на близост при произволно сдвояване с друг клъстер. Цели се сливане на най-близките по дистанция. След сливане на цялото ниво в дървото, се продължава към следващото – новоизградено. Тези стъпки се повтарят или до достигане на k или до конвергенция от един клъстер. Нека разстоянието D между обекти x_i и x_j се представи по следния начин:

$$D = \left|\overrightarrow{x_i} - \overrightarrow{x_j}\right| = \sqrt{\left|\overrightarrow{x_i} - \overrightarrow{x_j}\right|^2} = \sqrt{(|x_i|\langle x_i| - |x_j|\langle x_j|)(|x_i||x_i\rangle - |x_j||x_j\rangle)} \tag{34}$$

Необходим е спомагателен квантов бит, който в този случай се конструира като сплетено състояние между двата обекта. Чрез измерването на сплетеното състояние ще се разбере връзката между вероятността, Еквлидовото разстояние и векторното произведение. Знаейки разстоянието, може да се изгради критерий за сливане на два клъстера. Рекласификацията на центроидите, т.е. тяхното обновяване, може да се дефинира като:

$$\overrightarrow{c_i} = |c_r||c_r\rangle = \frac{1}{N}\sum_{i=1}^{N_r}|x_i||x_i\rangle, \tag{35}$$

където N_r е броят на обектите в клъстера r, а x_i е самият обект от клъстера r.

Спомагателният бит - ancilla, подобно да приложението в **3.3** се поставя между състоянията $\overrightarrow{\boldsymbol{x_i}}$ и $\overrightarrow{\boldsymbol{x_j}}$ и се изгражда сплетено състояние:

$$|\varphi\rangle = \frac{1}{\sqrt{2}}(|0\rangle_{anc}|x_i\rangle + |1\rangle_{anc}|x_j\rangle) \tag{36.1}$$

Конструкцията на сплетеното състояние може да се извърши при извикване от QRAM – този тип оперативна памет е разгледан в **Първа глава**. [Kong et al, 2017] изразява адресният регистър, съдържащ адрес на суперпозиция от $\sum_j p_j\,|j\rangle_a$, където j няма общо с x_j, p е някаква запазена вероятност, а a изразява спомагателен бит за връщане на резултат. Изходният регистър съдържа информационното суперпозиционно състояние $\sum_j p_j\,|j\rangle_a|D_j\rangle_d$, асоциирано с адресния регистър. D в случая представлява данните: $|D\rangle_d = |x_i\rangle + |x_j\rangle$. Приема се, че предварително гейт на Адамар е инициализирал състоянието в суперпозиция $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. За изходния резултат на данните се образува следното сплетено състояние:

$$f\left(\frac{|0\rangle}{\sqrt{2}}, \frac{|1\rangle}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}}(|0\rangle|x_i\rangle + |1\rangle|x_j\rangle) \tag{36.2}$$

Извършва се проекционно измерване на $|\varphi\rangle$, както се описва в **Първа глава**, и се проверява дали измерването се проектира върху $|\phi\rangle$:

$$|\phi\rangle = {(|x_i||0\rangle - |x_j||1\rangle)}\Big/{\sqrt{|x_i|^2 + |x_j|^2}} \quad (36.3)$$

С цел генерация на $|\phi\rangle$ се извърваш унитарна трансформация e^{-iHt} за времева еволюция за състояние $\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)\otimes|0\rangle$, където $H = (|\overrightarrow{x_i}||0\rangle\langle 0| + |\overrightarrow{x_j}|x_j\rangle)\otimes\sigma_x$. Съгласно (7.4) резултатът от операцията ще бъде:

$$\frac{1}{\sqrt{2}}[\cos(|\overrightarrow{x_i}|t)|0\rangle - \cos(|\overrightarrow{x_j}|t)|1\rangle]\otimes|0\rangle - \frac{i}{\sqrt{2}}[\sin(|\overrightarrow{x_i}|t)|0\rangle - \sin(|\overrightarrow{x_j}|t)|1\rangle]\otimes|1\rangle. \quad (36.4)$$

При спазване на условието за *t*, т.е. $|\overrightarrow{x_i}|t, |\overrightarrow{x_j}|t \ll 1$, то, следвайки (36.4), състоянието $|\phi\rangle$ ще бъде достигнато с вероятност от $\frac{1}{2}(|\overrightarrow{x_i}|^2 + |\overrightarrow{x_j}|^2)t^2$. При повторно изпълняване на проекционното измерване за $|\varphi\rangle$, се определя $|\phi\rangle$ с вероятност *p,* където:

$$p = p(|\phi\rangle) = \langle\varphi|{M^\dagger}_\phi M_\phi|\varphi\rangle \quad (36.5)$$

Знаейки, че $M_\phi = |\phi\rangle\langle\phi|$, то (36.5) може да се запише по следния начин:

$$p = p(|\phi\rangle) = \langle\varphi|\phi\rangle\langle\phi|\varphi\rangle =$$

$$\frac{\langle x_i|\langle 0| + \langle x_j|\langle 1|}{\sqrt{2}} \cdot \frac{|x_i||0\rangle - |x_j||1\rangle}{\sqrt{|x_i|^2 + |x_j|^2}} \cdot \frac{|x_i|\langle 0| - |x_j|\langle 1|}{\sqrt{|x_i|^2 + |x_j|^2}} \cdot \frac{|0\rangle|x_i\rangle + |1\rangle|x_j\rangle}{\sqrt{2}} =$$

$$\frac{(|x_i|\langle x_i| - |x_j|\langle x_j|).(|x_i||x_i\rangle - |x_j||x_j\rangle)}{2\left(|x_i|^2 + |x_j|^2\right)} \quad (36.6)$$

3.3.2 Откриване на сходства между елементи и получаване на центроиди

Комбинирани (34) и (36.6) водят до разстоянието, което може да се измери между $\overrightarrow{x_i}$ и $\overrightarrow{x_j}$:

$$D = \sqrt{2p\left(|x_i|^2 + |x_j|^2\right)} \quad (36.7)$$

При разкриване на скобите от (36.6) следва:

$p = p(|\phi\rangle) = \frac{|x_i|^2 - |x_i||x_j|\langle x_i|x_j\rangle - |x_i||x_j|\langle x_j|x_i\rangle + |x_j|^2}{2\left(|x_i|^2 + |x_j|^2\right)}$. Скаларното произведение на $|x_i\rangle$ и $|x_j\rangle$ се получава по следния начин:

$$\langle x_i|x_j\rangle + \langle x_j|x_i\rangle = {(1-2p)\left(|x_i|^2 + |x_j|^2\right)}\Big/{(|x_i||x_j|)} \quad (37)$$

Получаване на новите центроиди при N_s входни данни във всеки от s-тите клъстери се изразява чрез:

$$\frac{1}{N_s}\sum_{k=1}^{N_s}|x_k||x_k\rangle \tag{38}$$

3.3.3 Оценка на качеството на получените клъстери

След като вече е установен методът за клъстериране, може да се премине към следващ етап, а именно представяне на резултатите или качествена оценка на получените клъстери. Втората техника се използва рядко в системите, занимаващи се с динамични редове, защото потокът данни се приема за голям. Това означава, че дори и оптимизирани, алгоритмите за клъстериране остават изчислително сложни и системата не може лесно да се самоверифицира, особено когато времето за реакция е важно. Въпреки това, ако се приеме, че квантовото представяне на тези методологии би довело до значително ускорение, то се предполага и че стъпката за самооценяване, също може да бъде ускорена по подобен начин, като се запази като част от алгоритъма, тъй като реално няма да се оказва мястото, където е възможно да настъпи допълнително забавяне.

Друга важна роля при оценка на качеството на получените клъстери е конвергенция на алгоритъма, т.е. или той приключва до определен брой итерации, или докато не наблюдават промени в състава на клъстерите. Възможно е нито едно от двете условия да не е приемливо при динамичните редове, затова се налага трети вариант с използване на качествената оценка. При достатъчно приемливо качество на клъстерите, системата може да премине към представяне на резултатите и така да се избегне ненужно за конкретния експеримент итериране. Такъв подход бе проучен в [Andreev, 2018] и [Andreev et al, 2018], а именно т. нар. *методът Силует*[2]. Неговият начин на работа може да се дефинира просто чрез следната система:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \textit{if } a(i) < b(i) \\ 0, & \textit{if } a(i) = b(i), \\ b(i)/a(i) - 1, & \textit{if } a(i) > b(i) \end{cases} \tag{39}$$

където a е средно-статистическото разстояние от наблюдаемото i до точките от клъстера, на който принадлежи. b е минималното средно-статистическо разстояние на i спрямо точките от друг произволен клъстер. С други думи, b е средно-статистическото несъответствие между точка i и точките на най-близкия клъстер, до нейния клъстер. Метриките за разстояние бяха описани текста по-горе, както и още в **Първа глава**. От (6) и (7) може да се направи извода, че това отново е близост на два квантови бита, която бе наречена също *прецизност*. В конкретния случай, близостта между квантовото

[2] Бел. прев. от англ. Silhouette – силует.

състояние на наблюдавания и всички останали спрямо (21). За *n*-мерност на текущия за итерацията клъстер може да се използва следният израз:

$$a(i) = \frac{1}{n_c} \sum_{j \in c, j=1}^{n_c} \langle F(\rho, \rho_j{}') \rangle_{\rho_0}, \tag{40.1}$$

$$b(i) = min \frac{1}{s} \sum_{c=1}^{s} \frac{1}{n_s} \sum_{i \notin S_c, j=1}^{n_s} \langle F(\rho, \rho_j{}') \rangle_{\rho_0}, \tag{40.2}$$

като близостта на двете квантови състояния се маркира с: $\langle F \rangle = \langle F(\rho, \rho') \rangle_{\rho_0}$, където ρ и ρ′ са матриците за плътността на *i*-тата променлива, както и всички останали, съответно при първоначални състояния ρ_0. Допълнително, *s* е броят на клъстерите, а *c* бележи индекса на текущия клъстер. n_s е броят на елементи в клъстера *S,* за който *i* не принаделжи на него, т.е. S_c е в контекста на *b*:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)}, \text{ за } \forall i \in \vec{x} \tag{41}$$