ISSN 2658 – 7505
Выпуск №1, 2020
Электронный научный журнал «Вестник молодёжной науки России»

УДК 004.658

ВЕРОЯТНОСТНАЯ КЛАСТЕРИЗАЦИЯ МНОГОМЕРНЫХ ОБЪЕКТОВ С ИСПОЛЬЗОВАНИЕМ САМООРГАНИЗУЮЩЕЙСЯ НЕЙРОННОЙ СЕТИ

Кузнецова А.В. *alvitkuz@yandex.ru,* ***Алгазали С.М.М.*** *alghazali.salah@yandex.ru*
Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова,
г. Новочеркасск

В рамках разработки ядра интеллектуальной системы анализа производительности СУБД предложен комбинированный метод кластеризации SQL-запросов, представленных набором статистических параметров производительности. Значительные наборы многомерных векторов предварительно обрабатываются самоорганизующейся нейронной картой Кохонена. Полученная карта служит достаточно хорошим приближением для работы модифицированного EM-алгорима разделения гауссовой смеси распределений. Предложенный метод показал свою эффективность как при отсутствии априорной информации о количестве кластеров, так при наличии областей со слабой локализацией объектов.

Ключевые слова: SQL-запрос, самоорганизующаяся карта – SOM, модель смеси гауссовых распределений – GMM.

PROBABLE CLUSTERING OF MULTIDIMENSIONAL OBJECTS USING A SELF-ORGANIZING MAP

Kuznetsova A.V., Alghazali C.M.M.
Platov South-Russian State Polytechnic University (NPI), Novocherkassk

As part of the development of the core of an intelligent DBMS performance analysis system, a combined method for clustering SQL queries is presented, which are represented by a set of statistical performance parameters. Significant sets of multidimensional vectors are pre-processed by Kohonen's self-organizing neural map. The resulting map serves as a good enough approximation for the operation of the modified EM algorithm for separating a Gaussian mixture of distributions. The proposed method has shown its effectiveness both in the absence of a priori information on the number of clusters, and in the presence of areas with weak localization of objects.

Keywords: SQL query, self-organizing map – SOM, Gaussian Mixture Model – GMM.

Введение

Классические алгоритмы кластерного анализа осуществляют направленный поиск в сравнительно небольшом подмножестве пространства решений, используя различного рода априорные ограничения (на число кластеров или их форму, на порядок включения объектов в группы и т.д.). При этом нахождение строго-оптимального решения не гарантируется. Для поиска решения в случае отсутствии априорных сведений о количестве кластеров применяются более сложные методы, такие как генетические (эволюционные) алгоритмы, нейронные сети или алгоритмы с разными параметрами настройки, с различными наборами переменных. После построения ансамбля проводится нахождение итогового коллективного решения.

В последнее время различными авторами [1-3] предлагаются способы повышения устойчивости группировочных решений, основанные на применении коммбинации различных подходов и методов. Наибольший эффект такие решения дают в том случае, когда обработке подвергаются значительные (свыше 10^3) объёмы данных, в пространстве входных векторов которых имеют место не только

хорошо локализованные подпространства, но и разреженные области, содержащие малое число объектов. Одной из основных задач объединения нескольких методов типа является устранение тех или иных недостатков или использование результатов одного метода в качестве начального приближения другого.

Комбинация нейросетевого и вероятностного методов кластеризации

В рамках построения системы для анализа производительности SQL-запросов был разработан робастный нейросетевой алгоритм обучения самоорганизующейся карты Кохонена (англ. S*elf-Organizing Maps, SOM*), позволяющий визуализировать большие объёмы многомерных неразмеченных входных векторов, максимально сохраняющий их топологические взаимоотношения и минимизирующий случайные негативные влияния внешней среды [4, 5]. Последующий анализ карт, обученных на реальных данных недельных и месячных интервалов, показал, что, не смотря на введенное «загрубление», на отдельных участках имеет место эффект размытия, который затрудняет однозначное выделение кластеров. Использование алгоритма k-means и ему подобных для последующего выявления кластеров на объектах, предварительно соотнесенных с нейронами карты невозможно в силу вышеуказанного отсутствия предположений о возможном числе кластеров. Результаты классического алгоритма U-matrix очень сильно разнятся для одного и того набора данных в зависимости от выбора начальной точки обхода узлов сети. Для того чтобы алгоритм всегда выдавал практически однозначные результаты, области в U-матрице должны быть вогнутыми в пределах разрешения используемого квантования. Для сложных наборов данных, имеющих сильные разрывы плотности, высокие сгущения и значительное число выбросов, это требование выполняется далеко не всегда. Именно к таким наборам данных следует отнести исследуемые множества SQL-запросов [6].

Исходя из этого принято решение об использовании вероятностного метода обучения без учителя, который бы дополнил модель нейросетевой кластеризации объектов целевой программной системы. Таким наиболее известным и широко используемым в машинном обучении методом является Гауссова смесь распределений (англ. Gaussian Mixture Model, GMM). Это одна из форм оценки плотности распределения объектов в кластерах на основе предположения о том, что каждый используемый параметр объекта подчиняется нормальному закону. Большое число случайных факторов, оказывающих влияние на наборы статистических параметров SQL-объектов, а также анализ достаточно больших (>50) наборов данных, попавших в область притяжения отдельных нейронов SOM свидетельствуют в пользу принятого решения.

Формирование гауссовой смеси распределений в рассматриваемом методе строится на основе объединения уже выделенных подобластей сосредоточения объектов. Другими словами, в качестве входных объектов алгоритма выделения кластеров выступают активные нейроны $\mathbf{d}_j$, весовые коэффициенты $\mathbf{w}^{(dj)}$ которых являются «нейросетевой» регуляризционной оценкой подмножества принадлежащих этим нейронам элементов данных $\{\mathbf{x}_1^{(dj)}, \mathbf{x}_2^{(dj)}, \ldots, \mathbf{x}_n^{(dj)}\}$.

Каждая подобласть представлена «гиперэллипсоидом» с размерами осей, соответствующими среднеквадратическому отклонению параметров элементов

ISSN 2658 – 7505
Выпуск №1, 2020
Электронный научный журнал «Вестник молодёжной науки России»

данных от соответствующего весового коэффициента, и поворотом, соответствующим корреляциям различных параметров объектов, попавших в область притяжения $\mathbf{d}_j$. С формальной точки зрения имеется множество **D** предварительно сформированных n-мерных векторов $\mathbf{d}_j$, каждому из которых поставлено в соответствие непересекающееся подмножество $\mathbf{X}^{(\mathrm{d}j)}$ n-мерных векторов множества **X**:

$$\begin{aligned}&\mathbf{D}=\left\{\mathbf{d}_1,\mathbf{d}_2,...,\mathbf{d}_j,...,\mathbf{d}_{ND}\right\},\\&\forall\mathbf{d}_i\in\mathbf{D},\ \ \exists\mathbf{X}^{(\mathbf{d}_i)}\subset\mathbf{X}:\ \mathbf{X}=\left\{\mathbf{X}^{(\mathbf{d}_1)}\cup\mathbf{X}^{(\mathbf{d}_2)}\cup...\cup\mathbf{X}^{(\mathbf{d}_{ND})}\right\}.\end{aligned}\tag{1}$$

Необходимо получить модель H гауссовой смеси нормальных распределений с числом компонент $M << N_\mathrm{D}$, соответствующих естественной структуре плотности распределения объектов **X** в n-мерном пространстве признаков. Каждая компонента, в свою очередь, будет представлять собой n-мерный гауссиан с параметрами: **μ –** вектор средних, **Σ** – невырожденная матрица ковариации, β – единый параметр регуляризации, играющий роль масштабирующего коэффициента и регулирующий «размах» гауссианов совместно с диагональными элементами матрицы **Σ**, т.к. находится под знаком экспоненты:

$$P(\mathbf{x})=\mathcal{N}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)=\frac{1}{\sqrt{2\pi^n|\det\boldsymbol{\Sigma}|}}\exp\left(-\beta\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).\tag{2}$$

Исходя из (1) первоначальная плотность распределения вероятностей принадлежности множества объектов **X** множеству кластеров **D** представляет собой совокупность N_D гауссовых областей-кластеров согласно некоторой начальной гипотезе h_D вероятностной группировки данных:

$$P(\mathbf{X}\mid h_D)=\sum_{m=1}^{N_D}P(\mathbf{d}_m\mid h_D)P(m),\quad\sum_{m=1}^{N_D}P(m)=1,\quad P(m)\geq 0\tag{3}$$

$$h_D=\{\mathbf{w}_1,\mathbf{w}_2,...,\mathbf{w}_{N_D},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,...,\boldsymbol{\Sigma}_{N_D},P(1),P(2),...,P(N_D),\tag{4}$$

Гипотеза, описываемая набором параметров и априорной вероятностью $P(m)$ компонентов смеси, обладает предсказательной силой, позволяя оценить меру правдоподобия (Likelihood) имеющихся данных $P(\mathbf{X}|h)$ и предсказать вероятность распределения новых наборов данных. Оценить вероятность различных гипотез о способе разделения данных $P(h|\mathbf{X})$ можно на основе байесовского решающего правила:

$$P(h\mid\mathbf{X})=\frac{P(\mathbf{X}\mid h)P(h)}{P(\mathbf{X})}=\frac{P(\mathbf{X}\mid h)P(h)}{\sum_h P(\mathbf{X}\mid h)P(h)},\tag{5}$$

где апостериорная вероятность (Posterior) $P(h|\mathbf{X})$ уточняется на основе известной априорной вероятности гипотезы $P(h)$ (Prior), правдоподобия имеющихся данных $P(\mathbf{X}|h)$ и доказательства/свидетельства в пользу данной модели $P(\mathbf{X})$ (Evidence). Выбор наилучшей (наиболее вероятной) гипотезы в рамках модели H осуществляется на основе максимизации апостериорной вероятности $P(h|\mathbf{X})$:

$$h_{MPosterior}=\arg\max_h P(h\mid\mathbf{X}),\tag{6}$$

При отсутствии случайного и независимого множества прецедентов, содержащего метки кластеров $\{\mathbf{d}_\mathrm{i}, \mathrm{y_j}\}$, оптимальная гипотеза с числом кластеров M может быть определена на основе максимизации правдоподобия данных $P(\mathbf{X}|h)$:

ISSN 2658 – 7505
Выпуск №1, 2020
Электронный научный журнал «Вестник молодёжной науки России»

$$h_{MLikelihood} = \arg\max_{h} \ln P(\mathbf{X} \mid h) = \sum_{i=1}^{n} \ln \sum_{m=1}^{M} P(\mathbf{x}_i \mid m) P(m) . \quad (7)$$

В этом случае для её отыскания можно воспользоваться итерационным EM-алгоритмом (Expectation-Maximization), позволяющим избежать решения громоздкой оптимизационной задачи [7].

Модификации EM-алгоритма

Фундаментальный характер формулы Байеса позволяет единообразно сравнивать между собой гипотезы, выбирая наилучшую на основе максимизации апостериорной вероятности или на основе максимизации правдоподобия. Однако, такой подход имеет свои недостатки. С ростом числа проверяемых гипотез существенно увеличиваются затраты времени. При этом случайный выбор параметров начальных компонент для каждой генерируемой гипотезы – стандартная, но далеко не самая лучшая эвристика. Использование переборных алгоритмов на множестве **X** для начального размещения компонент требует определенных затрат времени и, опять-таки, не лишен элемента случайности. Но главное – встаёт вопрос об определении начальных априорных вероятностей для случайно заданных GMM.

Поскольку в качестве исходного числа компонент предлагается использовать число активных нейронов N_D, каждый из которых имеет своё подмножество объектов, начальные априорные вероятности компонент могут быть вычислены на основе количественных данных о размерах сформированных подмножеств. А поскольку тип модели кластеризации при отыскании наилучшего распределения меняться не будет, можно использовать модификацию EM-алгоритма, допускающую слияние источников. Для её реализации требуется дополнительный параметр алгоритма из формулы (2). Если на какой-то итерации t центры двух компонент совпадут $\mathbf{w}_i^t \approx \mathbf{w}_j^t$ с точностью β, то эти две компоненты объединятся в одну. При наличии нескольких пар претендентов на слияние, конфликтующих между собой, объединению подвергается та, которая обладает наименьшим расстоянием между центрами. Таким образом, с ростом числа итераций число изначально заданных областей N_D будет последовательно уменьшаться при каждом таком слиянии и реальное число кластеров M модели, будет существенно меньше первоначального.

В качестве критерия останова используется дивергенция или расстояние Кульбака-Лейблера [7], вычисляемое для двух последовательно полученных распределений:

$$D_{KL}(\mathcal{N}_t, \mathcal{N}_{t+1}) = \frac{1}{2}\left(tr(\boldsymbol{\Sigma}_{t+1}^{-1}\boldsymbol{\Sigma}_t) + (\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t)^{\mathrm{T}} \boldsymbol{\Sigma}_{t+1}^{-1}(\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t) - n + \ln\left(\frac{\det \boldsymbol{\Sigma}_{t+1}}{\det \boldsymbol{\Sigma}_t} \right) \right). \quad (8)$$

Отдельно следует остановиться на выборе параметра β, влияющем на радиус взаимодействия кластеров при их возможном слиянии наряду с начальной пространственной конфигурацией входных данных **D**. Параметр β в данном случае выполняет роль масштабируемого коэффициента и регуляризатора. Согласно численным экспериментам на тестовых данных в [5] установлено, что если радиус притяжения компонент, равный $\beta^{-1/2}$, превышает масштаб естественных сгущений, в которой эти компоненты расположены, то описанный выше EM-алгоритм приведет к их слиянию в одни кластер. Чем больше значение β, тем меньше радиус взаимодействия компонент и тем больше возможное число кластеров в модели. Малые значения дисперсии β, приводят к увеличению радиуса компонент вплоть

ISSN 2658 – 7505
Выпуск №1, 2020
Электронный научный журнал «Вестник молодёжной науки России»

до превышения масштаба разброса всех данных, что приводят к слиянию в один большой кластер.

Эксперименты с реальными данными показали, что в качестве значения β можно использовать величины, близкие к величине оптимальной топологической окрестности σ, определенной в [8]. Радиус притяжения компонент в этом случае гарантированно не превышает масштабы разброса всех данных и для значений σ_{opt} в диапазоне [7 .. 15] не выходит за границы [0,26 .. 0,38]. При этом обеспечивается хорошая сходимость алгоритма к M_{opt}=5..15 кластерам.

На рисунке 1. представлена обобщенная схема EM-алгоритма, разработанная с применением тензорного представления исследуемых объектов.

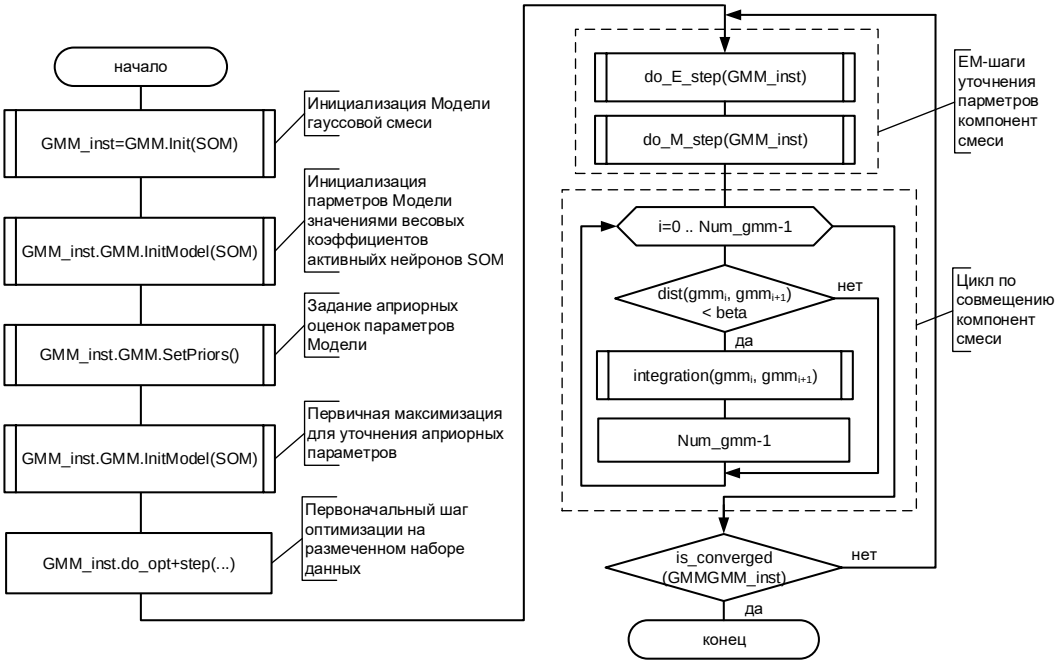

Рис. 1. Блок-схема модифицированного EM-алгоритма

Заключение

К достоинствам предложенного метода кластеризации можно отнести следующее:

– метод позволяет работать со сложными данными, разделяя этапы построения на сохранение глобальной структуры (SOM), а другой – на сохранение расстояний между объектами (GMM);

– построенная SOM даёт эффективный способ визуализации данных, в то время как использование результатов её обучения – весовых коэффициентов активных нейронов и начальное разделение объектов, принадлежащих этим нейронам, – даёт возможность начать вероятностное обучение модели кластеризации с достаточно удачных начальных условий, обеспечивая независимость от случайного начального положения центров каждой компоненты GMM.

– в процессе обучения вероятностной модели происходит постепенное её упрощение – сокращение избыточного количества компонент модели смесей, связанное с первоначальным большим размером карты SOM;

– оба этапа имеют относительную простоту и неплохую сходимость;
– полученные в результате разделения смеси распределений координаты $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ и апостериорные плотности распределения кластеров P(m) в дальнейшем могут использоваться для построения байесовского классификатора, способного обучаться на ограниченных подмножествах неэффективных SQL-запросов.

Список цитируемой литературы

1. Hazan, H., Saunders, D., Sanghavi, D.T., Siegelmann, H., Kozma, R. Unsupervised Learning with Self-Organizing Spiking Neural Networks // Proceedings of the International Joint Conference on Neural Networks, Vol. 2018-July, 10 October 2018, numb. 8489673.
2. Kamimura, R., Takeuchi H. Flexible Self-Organizing Maps by Information Maximization // Proceedings of the International Joint Conference on Neural Networks, Vol. 4, 2003, pp. 2734-2739.
3. Ultsch A., Morchen F. E. SOM-Maps tools for clustering, visualization data // Data Bionics Research Group / University of Marburg. Marburg. Germany. March 17. 2005.
4. Алгазали С.М.М., Айвазов В.Г., Кузнецова А.В. Модификация алгоритма обучения самоорганизующейся нейронной сети на основе механизма регуляризации. //Применение и развитие нейросетевого моделирования для решения фундаментальных задач в науке и технике. Материалы Международной молодежной научно-практической конференции 3– 4 декабря 2018 года г. Новочеркасск: Лик 2018., С. 62-65.
5. Кузнецова А.В., Алгазали С.М.М., Айвазов В.Г. Программа для ЭВМ: «Модифицированная самоорганизующаяся нейронная сеть на основе тензорных вычислений» (TZ-SOM). Свидетельство о государственной регистрации программ №201866479. Дата государственной регистрации в Реестре программ для ЭВМ 22 ноября 2018 г.
6. Алгазали С.М.М., Айвазов В.Г. Кузнецова А.В. Формирование модели SQL-запроса на основе статистических параметров производительности // Электронный научный журнал «Вестник молодёжной науки России» Выпуск №3, 2019
7. Прохоров Ю.В. Вероятность и математическая статистика: Энциклопедия: М.: Большая Российская энциклопедия – 1999. – 910 с.
8. Кузнецова А.В., Алгазали С.М.М., Айвазов В.Г. Особенности кластеризации SQL-запросов на основе механизма регуляризации // Известия вузов Сев-Кав. Регион. Технические науки, т.№3, Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова. – Новочеркасск, 4 октября 2019 г.

© А. В. Кузнецова, С.М.М. Алгазали, 2020