Лекция 9 Сети векторного квантования (LVQ). Современные модели и методы вычислений

Дисциплина: «Разработка алгоритмов для реализации методов машинного обучения(лек)»
гр:M094-6112-21-ауд:404
Кинтонова А.Ж.

Векторное квантование

В предыдущих разделах мы рассмотрели квантование выходного сигнала непрерывного источника для случая, когда квантование выполняется последовательно по отдельным отсчётам, т.е. скалярное квантование. В этом разделе мы рассмотрим совместное квантование блока символьных отсчётов или блока сигнальных параметров. Этот вид квантования называется **блоковым** или **векторным квантованием**. Оно широко используется при кодировании речи в цифровых сотовых системах связи.

Фундаментальный результат теории искажения заключается в том, что лучшую характеристику можно достичь векторным, а не скалярным квантованием, даже если непрерывный источник без памяти. Если, кроме того, отсчёты сигнала или параметры сигнала статистически зависимы, мы можем использовать зависимость посредством совместного квантования блоков отсчётов или параметров и таким образом достичь большей эффективности (более низкой битовой скорости) по сравнению с той, которая достигается скалярным квантованием.

Проблему векторного квантования можно сформулировать так. Имеем n-мерный вектор $X=\{x_1, x_2 \dots x_n\}$ с n вещественными, непрерывными амплитудами компонент $\{x_k, 1 \le k \le n\}$, которые описываются СФПВ $p(x_1, x_2 \dots x_n)$. Путём квантования вектор X превращается в другой n-мерный вектор $\tilde{X}$ с компонентами $\{\tilde{x}_k, 1 \le k \le n\}$. Выразим операции квантования оператором $Q()$, так что

$$\tilde{X} = Q(X), \qquad (3.4.31)$$

где $\tilde{X}$ - выход квантователя, когда на вход поступает вектор X.

В принципе векторное квантование блоков данных можно рассматривать как проблему распознавания образов, включающую в себя классификацию блоков данных через дискретное количество категорий или ячеек в соответствии с некоторым критерием точности, таким, например, как среднеквадратическая погрешность. Для примера рассмотрим квантование двумерных векторов $X=[x_1,x_2]$. Двумерное пространство разделяют на ячейки, как показано на рис. 3.4.3, где мы имеем произвольно выбранные шестиугольные ячейки $\{C_k\}$. Все входные векторы, которые попадают в ячейку C_k, квантуются в вектор $\tilde{X}_k$, который на рис. 3.4.3 отмечен как центр шестиугольника. В нашем примере иллюстрируются $L=37$ векторов, один для каждой из 37 ячеек, на которые разбито двумерное пространство. Обозначим ряд возможных выходных векторов как $\{\tilde{X}_k, 1\le k\le L\}$.

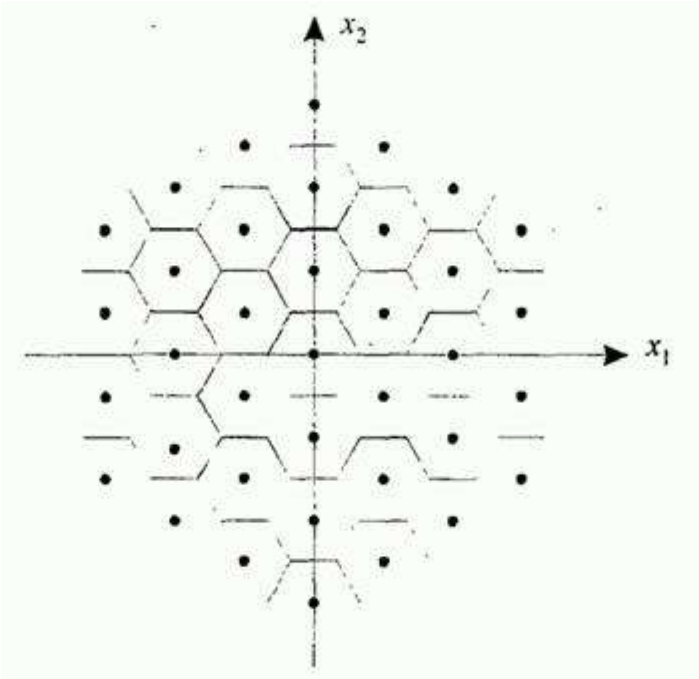

Рис. 3.4.3. Пример квантания в двухмерном пространстве

В общем, квантование n-мерного вектора X в n-мерный вектор $\tilde{X}$ ведёт к ошибке квантования или искажению $d(X,\tilde{X})$. Среднее искажение по ряду входных векторов X равно

$$D=\sum_{k=1}^{L}P(X\in C_k)E\left[d(X,\tilde{X}_k)\,|\,X\in C_k\right]=\sum_{k=1}^{L}P(X\in C_k)\int_{X\in C_k}d(X,\tilde{X}_k)p(X)dX \quad , \qquad (3.4.32)$$

где $P(X\in C_k)$ - вероятность того, что вектор X попадёт в ячейку C_k, а $p(X)$ - СФПВ n случайных величин. Как и в случае скалярного квантования, мы можем минимизировать D путём выбора ячеек $\{C_k, 1<k\le L\}$ при заданной

ФПВ $p(X)$. Обычно используемая мера искажений - среднеквадратическая ошибка (l_2 - норма) определяется как

$$d_2(X,\tilde{X}) = \frac{1}{n}(X-\tilde{X})^T(X-\tilde{X}) = \frac{1}{n}\sum_{k=1}^{n}(x_k-\tilde{x}_k)^2 \qquad (3.4.33)$$

или, в более общем виде, взвешенная среднеквадратическая ошибка

$$d_{2W}(X,\tilde{X}) = (X-\tilde{X})^T W(X-\tilde{X}), \qquad (3.4.34)$$

где W - положительно определённая взвешивающая матрица. Обычно мера W выбирается как обратная по отношению к матрице ковариаций входных данных X.

Другая мера искажений, которая иногда используется, является частным случаем l_p нормы и определяется как

$$d_p(X,\tilde{X}) = \frac{1}{n}\sum_{k=1}^{n}\left|x_k-\tilde{x}_k\right|^p. \qquad (3.4.35)$$

Частный случай, когда $p=1$, часто используется как альтернатива случаю $p=2$.

Векторное квантование не ограничивается квантованием блока сигнальных отсчётов источника сигнала. Его можно использовать для квантования ряда параметров, извлечённых из данных. Например, при линейном кодировании с предсказанием (ЛКП), описанном в разделе 3.5.3, параметры, извлечённые из сигнала, являются коэффициентами предсказания, которые являются коэффициентами для всеполюсной фильтровой модели источника, который генерирует наблюдаемые данные. Эти параметры можно рассматривать как блок и квантовать как блок символов, используя некоторую подходящую меру искажений. В случае кодирования речи подходящей мерой искажений, которую предложили Итакура и Саити (1986, 1975), является взвешенная среднеквадратическая ошибка, где взвешивающая матрица W выбрана как нормированная матрица автоковариации Φ наблюдаемых данных.

При кодировании речи альтернативным рядом параметров, которые могут быть квантованы как блок и переданы к приёмнику, могут быть коэффициенты отражения (см. ниже) $\{a_{ii}, 1 \le i \le m\}$.

Еще один ряд параметров, которые иногда используются для векторного квантования при линейном кодировании с предсказанием речи, содержит логарифмические отношения $\{r_k\}$, которые выражаются через коэффициенты отражения

$$r_k = 1\log\frac{1+a_{kk}}{1-a_{kk}},\quad 1 \le k \le m. \qquad (3.4.36)$$

Теперь вернемся к математической формулировке векторного квантования и рассмотрим разбиение n-мерного пространства на L ячеек $\{C_k, 1 < k \le L\}$ с точки зрения минимизации среднего искажения по всем L-уровневым квантователям. Имеется два условия для минимизации. Первое заключается в том, что оптимальный квантователь использует селекцию по правилу ближайшего соседа, которое можно выразить математически как

$$Q(X) = \tilde{X}_k,$$

если, и только если

$$D(X,\tilde{X}_k) \le D(X,\tilde{X}_j),\quad k \ne j,\quad 1 \le j \le L. \qquad (3.4.37)$$

Второе условие, необходимое для оптимизации, заключается в том, что каждый выходной вектор $\tilde{X}_k$ выбирается так, чтобы минимизировать среднее искажение в ячейке C_k. Другими словами, $\tilde{X}_k$ - это вектор в C_k, который минимизирует

$$D_k = E\left[d(X,\tilde{X}) \mid X \in C_k\right] = \int_{X\in C_k} d(X,\tilde{X})p(X)dX. \qquad (3.4.38)$$

Вектор $\tilde{X}_k$, который минимизирует D_k, назван **центроидом** ячейки.

Таким образом, эти условия оптимизации определяют разбиение n-мерного пространства на ячейки $\{C_k, 1 \le k \le L\}$, когда СФПВ $p(X)$ известна. Ясно, что указанные два условия обобщают задачу оптимального квантования скалярной величины оптимизации на случай квантования n-мерного вектора. В общем, мы ожидаем, что кодовые векторы более тесно группируются в областях, где СФПВ $p(X)$ велика, и, наоборот, разрежены в областях, где $p(X)$ мала.

В качестве верхней границы искажений векторного квантования мы можем использовать величину искажений оптимального скалярного квантователя, и эту границу можно применить для каждой компоненты вектора, как было описано в предыдущем разделе. С другой стороны, наилучшие характеристики, которые могут быть достигнуты оптимальным векторным квантователем, определяются функцией скорость-искажение или, что эквивалентно, функцией искажение-скорость. Функция искажение-скорость, которая была введена в предыдущем разделе, может быть определена в контексте
векторного квантования следующим образом. Предположим, мы формируем вектор X размерности n из n последовательных отсчётов $\{x_k\}$. Вектор X квантуется в форму $\tilde{X} = Q(X)$, где $\tilde{X}$ - образованный

рядом $\{\tilde{X}_m, 1 < m \le L\}$. Как было описано выше, среднее искажение D, получаемое при представлении X через $\tilde{X}$, равно $E\left[d(X,\tilde{X})\right]$, где $d(X,\tilde{X})$ - это искажение на одно измерение. Например,

$$d(X,\tilde{X}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \tilde{x}_k)^2 .$$

Минимально достижимая средняя битовая скорость, с которой могут быть переданы векторы $\{\tilde{X}_m, 1 < m \le L\}$, равна

$$R = \frac{H(\tilde{X})}{n} \text{ бит/отсчет,} \qquad (3.4.39)$$

где $H(\tilde{X})$ - энтропия квантованного выхода источника, определяемая как

$$H(\tilde{X}) = -\sum_{k=1}^{L} p(\tilde{X}_i)\log_2 P(\tilde{X}_i) . \qquad (3.4.40)$$

Для данной средней скорости R минимально достижимое искажение

$$D_n(R) = \min_{Q(X)} E\left[d(X,\tilde{X})\right], \qquad (3.4.41)$$

где $R \ge H(\tilde{X})/n$ и минимум в (3.4.41) берётся по всем возможным отображениям $Q(X)$. В пределе, когда размерность n стремится к бесконечности, получаем

$$D(R) = \lim_{n\to\infty} D_n(R), \qquad (3.4.42)$$

где $D(R)$ - это функция искажение-скорость, которая была введена в предыдущем разделе. Из этого изложения очевидно, что функция искажение-скорость может быть как угодно приближена к пределу путём увеличения размерности n векторов.

Изложенный выше подход приемлем в предположении, что СФПВ $p(X)$ вектора данных известна. Однако на практике СФПВ $p(X)$ данных может быть неизвестна. В этом случае, возможно адаптивно выбрать квантованные выходные векторы с использованием ряда обучающих векторов $X(m)$. Конкретнее, предположим, что мы имеем ряд из M векторов, причём M намного больше, чем L $(M \gg L)$. Итеративный групповой алгоритм, названный **алгоритмом K средних**, где в нашем случае $K = L$, может быть применён к обучающим векторам. Этот алгоритм итеративно делит M обучающих векторов на L групп так, что два необходимых условия оптимальности выполняются. Алгоритм K средних может быть описан так, как дано ниже [Макхоул и др. (1985)].

Алгоритм K средних

Шаг 1. Инициализируется начальный номер итерации $i = 0$. Выбирается ряд выходных векторов $\tilde{X}_k(0)$, $1 \le x \le L$.

Шаг **2.** Обучающие векторы $\{X(m), 1 < m \le M\}$ классифицируются в группы $\{C_k\}$ посредством правила ближайшего соседа:

$X \in C_k(i)$ если $D(X, \tilde{X}_k(i)) \le D(X, \tilde{X}_j(i))$ для всех $k \ne j$.

Шаг 3. Пересчитываются (для $(i+1)$-го шага) выходные векторы каждой группы путём вычисления центроида

$$\tilde{X}_k(i) = \frac{1}{M_k} \sum_{X \in C_k} X(m), \qquad 1 \le k \le L,$$

для обучающих векторов, которые попадают в каждую группу.

Кроме того, рассчитывается результирующее искажение $D(i)$ на i-й итерации.

Шаг 4. Заканчивается тестирование, если $D(i-1) - D(i)$ относительно мало. В противном случае следует идти к шагу 2.

Алгоритм K средних приводит к локальному минимуму (см. Андерберг, 1973; Линде и др., 1980). Начиная этот алгоритм различными рядами начальных выходных векторов $\{X_k(0)\}$ и каждый раз выполняя оптимизацию, описанную алгоритмом K средних, можно найти глобальный оптимум. Однако вычислительные затраты этой поисковой процедуры могут ограничить поиск немногими инициализациями.

Если мы один раз выбрали выходные векторы $\{\tilde{X}_k, 1 < k \le L\}$, каждый сигнальный вектор $X(m)$ квантуется в выходной вектор, который является ближайшим к нему с точки зрения выбранной меры искажения. Если вычисление включает в себя оценку расстояния между $X(m)$ и каждым из L возможных выходных векторов $\{\tilde{X}_k\}$, процедура образует полный поиск. Если предположим, что каждое вычисление требует n умножений и сложений, то общее требуемое число вычислений для полного поиска равно

$$\wp = nL \tag{3.4.43}$$

умножений и сложений на входной вектор.

Если мы выбрали L как степень 2, то $\log_2 L$ определяет число бит, требуемых для представления каждого вектора. Теперь, если R обозначает битовую скорость на отсчёт [на компоненту или на измерение $X(m)$], имеем $nR = \log_2 L$ и, следовательно, вычислительные затраты

$$\wp = n2^{nR}. \tag{3.4.44}$$

Заметим, что число вычислений растёт экспоненциально с параметром размерности n и битовой скорости R на измерение. Вследствие этого экспоненциального роста вычислительных затрат векторное квантование применяется в низкобитовых кодерах источника, таких как кодирование коэффициентов отражения или логарифмических отношений в линейном кодировании речи с предсказанием.

Вычислительные затраты, связанные с полным поиском, можно уменьшить при помощи изящного субоптимального алгоритма (см. Чанг и др., 1984; Гершо, 1982).

Чтобы продемонстрировать пользу векторного квантования по сравнению со скалярным квантованием, мы представим следующий пример, взятый у Макхоула и др. (1985).

Пример 3.4.1. Пусть X_1 и X_2 являются двумя случайными величинами с равномерной СФПВ:

$$p(x_1, x_2) = p(X) = \begin{cases} \dfrac{1}{ab} & (X \in C), \\ 0 & (\text{для других } X), \end{cases} \tag{3.4.45}$$

где C - прямоугольная область, показанная на рис. 3.4.4. Заметим, что прямоугольник повёрнут на 45° относительно горизонтальной оси. На рис. 3.4.4 показаны также собственные плотности вероятности $p(x_1)$ и $p(x_2)$.

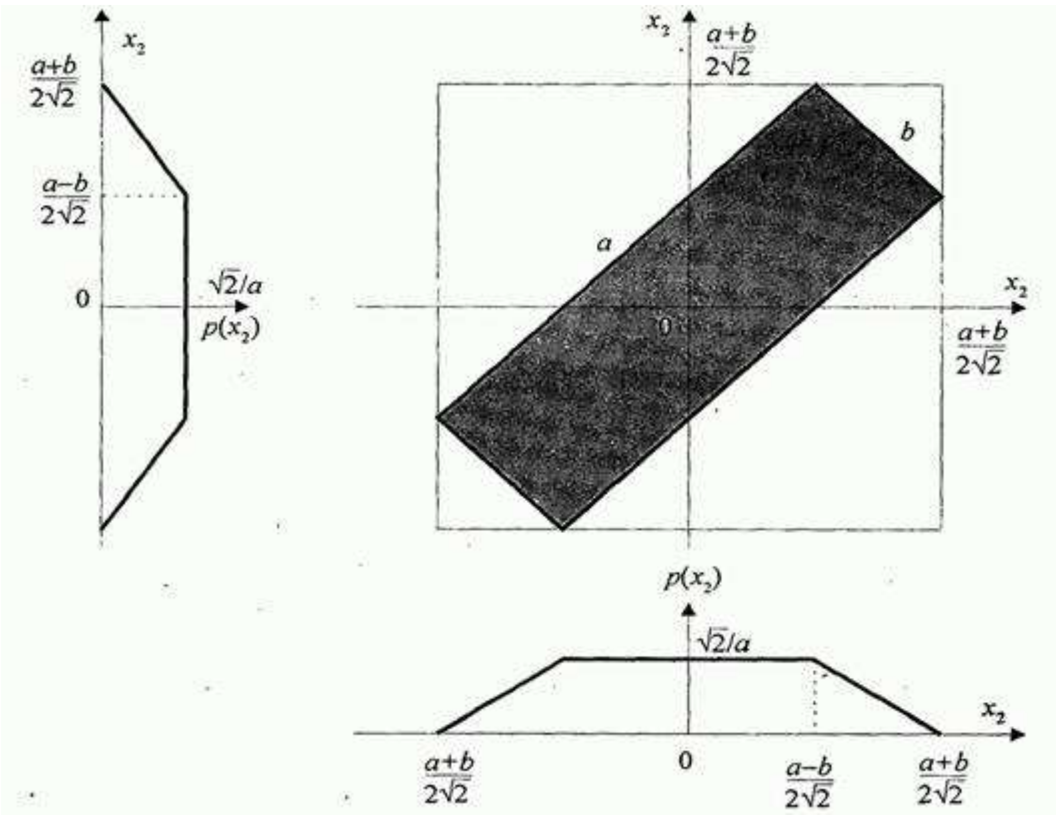

Рис. 3.4.4. Равномерная ФПВ в двух измерениях (Макхоул и др., 1985)

Если мы квантуем x_1 и x_2 раздельно, используя одинаковые интервалы квантования длины Δ, то требуемое число уровней квантования

$$L_1 = L_2 = \frac{a+b}{\sqrt{2}\Delta}. \qquad (3.4.46)$$

Следовательно, для кодирования вектора $X = [x_1 x_2]$ потребуется число бит

$$R_x = R_1 + R_2 = \log_2 L_1 + \log_2 L_2,$$

$$R_x = \log_2 \frac{(a+b)^2}{2\Delta^2}. \qquad (3.4.47)$$

Таким образом, скалярное квантование каждой компоненты эквивалентно векторному квантованию с общим числом уровней

$$L_x = L_1 L_2 = \frac{(a+b)^2}{2\Delta^2}. \qquad (3.4.48)$$

Видим, что это приближение эквивалентно покрытию большой площади, которая охватывает прямоугольник посредством квадратных ячеек, причём каждая ячейка представляет одну из L_x областей квантования.

Поскольку $p(X)=0$, за исключением $X \in C$, такое кодирование является расточительным и приводит к увеличению битовой скорости.

Если же мы покроем только область, где $p(X) \neq 0$, квадратиками, имеющими площадь Δ^2, то общее число уровней, которые образуются, определяется площадью прямоугольника, делённой на Δ^2, т.е.

$$L'_x = \frac{ab}{\Delta^2}. \qquad (3.4.49)$$

Следовательно, разница в битовой скорости при скалярном и векторном методах квантования равна

$$R_x - R'_x = \log_2 \frac{(a+b)^2}{2ab}. \qquad (3.4.50)$$

Для случая, когда $a = 4b$, разница в битовой скорости

$$R_x - R'_x = 1,64 \text{ бит/вектор.}$$

Следовательно, векторное квантование на 0,82 бит/отсчёт лучше, чем скалярное, при тех же искажениях.

Интересно заметить, что линейное преобразование (поворот на 45º) декоррелирует X_1 и X_2 и делает две случайные величины статистически независимыми. Тогда скалярное квантование и векторное квантование достигают одинаковой эффективности.
Хотя линейное преобразование может декоррелировать вектор случайных величин, оно не приводит к статистически независимым случайным величинам в общем случае. Следовательно, Векторное квантование будет всегда равняться или превосходить по характеристикам скалярный квантователь (см. задачу 3.40).

Векторное квантование применяется при различных методах кодирования речи, включая сигнальные методы и методы базовых моделей, которые рассматриваются в разд. 3.5. В методах, основанных на базовых моделях, таких как линейное кодирование с предсказанием, векторное квантование делает возможным кодирование речи на скоростях ниже 1000 бит/с (см. Бузо и др., 1980; Роукос и др., 1982; Пауль, 1983). Если использовать методы кодирования сигналов, возможно получить хорошее качество речи на скоростях передачи 16 000 бит/с, что эквивалентно скорости кодирования $R=2$ бит/отсчёт. За счёт дополнительных вычислительных усложнений в будущем станет возможным использовать сигнальные кодеры, обеспечивающие хорошее качество речи при скорости кодирования $R=1$ бит/отсчёт.