СРАВНЕНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ ДИКТОРА

Любимов Николай, Михеев Евгений, Лукин Алексей
Московский Государственный Университет им. М. В. Ломоносова

1. Введение

В этой статье мы рассмотрим задачу текстонезависимой идентификации дикторов. Один из наиболее современных подходов к решению этой задачи – использование *гауссовых смесей* (GMM) вида $f(x\,|\,\theta)=\sum_{k=1}^{K}\frac{\alpha_k}{\sqrt{\det(2\pi\Sigma)}}\exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right\}$ [1] для моделирования распределения таких характеристик диктора как мел-кепстральные коэффициенты (MFCC) [1] или кепстральные коэффициенты линейного предсказания (LPCC) [2]. Классификация достигается выбором класса диктора с максимальным правдоподобием на заданном участке данных. Более сложный подход [3] использует дикриминативные методы (например, метод опорных векторов) для разделения акустических классов. Также существуют гибридные системы [4], комбинирующие метод опорных векторов и GMM.

Мы рассмотрим простейшую систему идентификации диктора, в которой можно выделить 3 основных этапа: 1) предобработка на основе MFCC и использования детектора речи, 2) начальная кластеризация в пространстве признаков, 3) переоценка параметров гауссовых смесей на основе EM-алгоритма (Expectation Maximization) [5]. Решающее правило в задаче идентификации формулируется в виде принципа максимального правдоподобия модели диктора на наборе входных векторов признаков X:

$$i^* = \arg\max_{i=1,2,\ldots,N}\sum_{x\in X}\log f(x|\theta_i), \qquad (1)$$

где $\theta_i=\left(\left\{\alpha_k^{(i)},\mu_k^{(i)},\Sigma_k^{(i)}\right\}_{k=1}^{K}\right)$, $i=1,2,\ldots,N$ – набор параметров гауссовых смесей, а каждый i-ый набор определяет модель диктора, заявленного на поиск.

В данной работе мы подробно остановимся на выборе способа начальной кластеризации для построения модели диктора. Ниже будут рассмотрены несколько известных алгоритмов кластеризации, использующих как четкую, так и нечеткую логику. Используя эти алгоритмы, мы ищем метод машинного обучения, на основе которого строятся модели с наименьшей ошибкой идентификации по формуле (1). Также рассматривается влияние детерминированности начального приближения EM-алгоритма на эффективность построенных моделей в задаче идентификации диктора. В конце статьи указаны некоторые перспективные направления исследования задачи начальной кластеризации в рамках акустического анализа речи.

2. Алгоритмы кластеризации

2.1. Алгоритм К-средних

К-средних – один из наиболее популярных алгоритмов кластеризации. Его основные достоинства – простота реализации и низкая вычислительная сложность [6]. Работая на дискретном наборе данных, алгоритм минимизирует расстояние между k центрами кластеров и точками исходных данных в соответствующем пространстве.

2.2. Алгоритм К-средних++

К-средних++ – модификация К-средних, отличающаяся инициализацией, которая рекурсивно инициализирует центры кластеров, на основании вероятности $\frac{D(x')^2}{\sum_{x \in X} D(x)^2}$, где $D(x)$ – кратчайшее евклидово расстояние между точкой x и ближайшим к ней уже выбранным центром [7].

Если набор центров C построен при помощи К-средних++, то потенциальная функция $V = \sum_{i=1}^{K} \sum_{x_j \in c_i} \left\| x_j - \mu_i \right\|$ удовлетворяет $E[V] \le 8(\ln k + 2) V_{opt}$.

2.3. Алгоритм Linde-Buzo-Gray (LBG)

LBG изначально представлен в [8]. Он очень похож на К-средних, за исключением того, что он обходит недетерминированность выбора начальных точек. Основная идея алгоритма – установить начальные центры в соответствии с главными компонентами входного вектора. Сначала находится среднее всего вектора. Затем область входных данных разбивается на 2 кластера по оси главной компоненты. Далее с помощью стандартного К-средних вычисляются 2 кластера. Затем берется кластер с бо́льшим радиусом и снова делится пополам. Так продолжается до достижения нужного количества кластеров.

2.4. Алгоритм Fuzzy C-means (FCM)

FCM – один из наиболее популярных алгоритмов нечёткой кластеризации. Он делит область данных на K сферических кластеров. Основная идея алгоритма – построение матрицы разбиения $\mathbf{U} = [u_{kn}]$, $k = 1,...,K$, $n = 1,2,...,N$, значениями которой являются вероятности принадлежности k-ому кластеру точки с индексом n [9]. На каждой итерации вычисляются центры кластеров $c_k = \langle x_n \rangle_{u_k}$ и пересчитывается матрица разбиения $u_{kn} = \left(d_{kn} \sum_{j=1}^{K} d_{jn}^{-1} \right)^{-1}$, где $d_{kn} = \left\| x_n - c_k \right\|$ – евклидова норма.

2.5. Алгоритм Гюстафсона-Кесселя (ГК)

Алгоритм ГК рассматривается как улучшение FCM [10]. Его основное отличие от FCM – введение для каждого кластера матрицы ковариации, вычисляемой по формуле $F_k = \langle (x_n - c_k)(x_n - c_k)^T \rangle_{u_k}$. На основании этой матрицы пересчитывается расстояние от точки до центра кластера в формуле: $d(x_n, c_k) = (x_n - c_k)^T \left(\det(F_k)^{\frac{1}{1+\dim}} F_k^{-1} \right) (x_n - c_k)$. Далее матрица разбиения вычисляется так же, как в алгоритме FCM.

3. Эксперимент

Для эксперимента мы выбрали базу русскоговорящих дикторов, записанную с телефонным качеством 8 кГц и частотным диапазоном 300–3400 Гц. Обучающий набор состоит из 40 дикторов, включающих мужские и женские голоса. Каждая запись, длиной в среднем 40 секунд, содержит фоновый шум, тишину и прочие неречевые данные. Тестовая база состоит из 10-, 20- и 30-секундных речевых фрагментов, причем каждый диктор представлен в среднем 5 записями. Результаты были получены независимо для каждой длины тестового файла, а затем скомбинированы для получения итогового результата.

4. Результаты

Точность распознавания нашей системы с различными методами инициализации показана на схеме 1. Шкала схемы показывает процентное соотношение корректно распознанных дикторов.

Другой интересный результат, полученный нами – зависимость точности распознавания от детерминированности метода инициализации. Для недетерминированных алгоритмов эксперимент был проведен 15 раз с различными начальными значениями, а затем получен усредненный результат. Как это видно из схемы 2, детерминированные методы дают примерно 1,5% прирост точности идентификации.

5. Выводы

Мы провели сравнение различных методов кластеризации для задачи идентификации диктора. Были рассмотрены и протестированы следующие алгоритмы: К-средних, К-средних++, Linde-Buzo-Gray, Fuzzy C-means и алгоритм Гюстафсона-Кесселя. Было установлено, что производительность модели, основанной на гауссовых смесях, зависит от детерминированности метода инициализации EM. Linde-Buzo-Gray лидирует среди алгоритмов четкой кластеризации, поскольку центры кластеров находятся в соответствии с главными компонентами, а не случайным образом, как в К-средних и К-средних++. Нечеткие алгоритмы показывают лучшие результаты, поскольку они более детерминированны и используют всю область для пересчета центров. Все алгоритмы, кроме Гюстафсона-Кесселя, ищут сферические кластеры, а ГК ищет эллиптические, что положительно отражается на его результатах.

6. Перспективы дальнейших исследований

Любую произносимую речь можно представить как непрерывную по времени траекторию в признаковом пространстве, однозначно задающую характеристики речевого тракта в момент произношения. Признаковое пространство можно разбить на подобласти таким образом, чтобы все траектории, соответствующие произношению конкретного слова, проходили через одни и те же подобласти. Это часто применяется фонетистами при построении диаграмм гласных [12]. Каждая подобласть определяет фонетическую единицу, последовательность которых задает фонетическую транскрипцию слова в естественном языке. В задаче идентификации диктора по голосу вопрос о качественном разбиении признакового пространства на подобласти проистекает из задачи более детального акустического анализа, включающего: 1) сравнение характера произношения одной фонемы различными

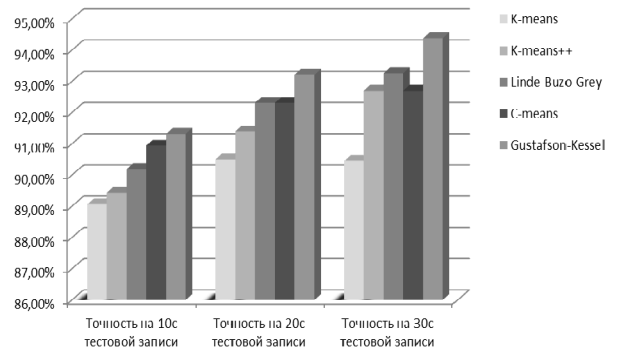

Схема 1. Сравнение различных техник кластеризации

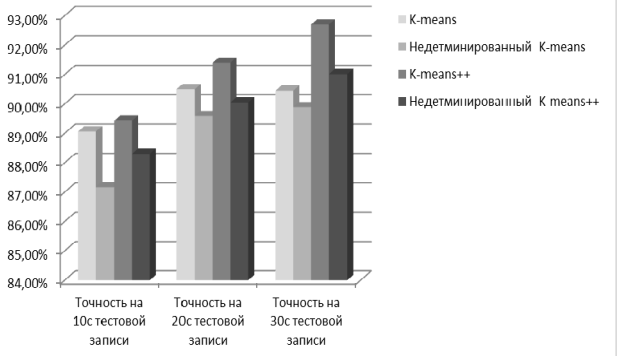

Схема 2. Сравнение детерминированных и недетерминированных алгоритмов

дикторами, 2) сравнения скоростей изменения фонем (темпа речи). В рассмотренной выше системе идентификации диктора соответствующего анализа произвести нельзя, так как правдоподобие вычисляется сразу для всех входных данных.

Дальнейшее исследование будет направлено на поиск оптимального автоматического разбиения исходного пространства акустических признаков на подобласти, где под оптимальностью понимается минимизация ошибки включения данных, полученных из различных фонем, в один кластер. Решение данной задачи позволит:

- значительно облегчить процесс построения систем автоматического распознавания речи за счет автоматизации затратной операции ручного транскрибирования речевых баз;
- осуществить возможность более детального акустического анализа голоса диктора за счет выделения схожих подобластей признакового пространства у различных дикторов;
- улучшить точность существующих систем верификации голоса на основе ключевых слов и фраз, использующих эргодические скрытые марковские цепи [13][14].

7. Список литературы

[1] D.A. Reynolds, R.C. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models,” in *Proc. IEEE ICASSP*, Vol. 3, No. 1, pp. 72–83, January 1995.

[2] W.-C. Chen, C.-T. Hsieh, E. Lai, “Multiband approach to Robust Text-Independent Speaker Identification”, in *IJCLCLP*, Vol. 9, No. 2, pp. 63–76, ACLCLP, August 2004.

[3] V. Wan, W.M. Campbell, “Support vector machines for speaker verification and identification,”, in *Proc. IEEE NNSPX’00*, Vol. 2, pp. 775–784, 2000.

[4] S. Fine, J. Navratil, R.A. Gopinath “A Hybrid GMM/SVM Approach to Speaker Identification,”, in *Proc. IEEE ICASSP’01*, Vol. 1, pp. 417–420, Salt Lake City, USA, 2001.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin “Maximum likelihood from Incomplete Data via the EM Algorithm”, *Journal of the Royal Statistical Society, Series B,* Vol. 39, No. 1, pp. 1–38, 1997.

[6] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” in *Proc. 5th Berkeley Symp. On Math. Stat. and Prob.,* Vol. 1, pp. 281–297, 1967

[7] D. Arthur, S. Vassilvitskii, “k-means++: The Advantages of Careful Seeding”, presented in *SODA*, 2007.

[8] Y. Linde, A. Buzo, R. Gray, “An Algorithm for Vector Quantizer Desing,” in *IEEE Trans. on Communications*, Vol. 28, pp. 84–94, 1980.

[9] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[10] D.E. Gustafson, W.C. Kessel, “Fuzzy Clustering with a Fuzzy Covariance Matrix,” in *Proc. of IEEE CDC*, 1979.

[11] J.P. Campbell, “Speaker Recognition: A Tutorial”, in *Proc. of the IEEE*, Vol. 85, No. 9, pp. 1437–1462, 1997.

[12]

[13] H. Lee, S. Chang, D. Yook, “A Voice Trigger System using Keyword and Speaker Recognition for Mobile Devices”, in *IEEE Trans. on Consumer Electronics*, Vol. 55, Issue 4, pp. 2377–2384, 2009.

[14] P. Li, J. Liang, B. Xu, “A Novel Instance Matching Based Unsupervised Keyword Spotting System*”,* in *2nd Int. Conf. on Innovative Computing, Information and Control (ICICIC’07)*, pp. 550–555, 2007.

Работа выполнена при поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы.