ANÁLISIS ARQUETÍPICO PARA APRENDIZAJE AUTOMÁTICO

Morten Mørup y Lars Kai Hansen

Grupo de Sistemas Cognitivos, Universidad Técnica de Dinamarca
Richard Petersens Plads, edificio 321, 2800 Lyngby, Dinamarca. Correo: {mm, lkh}@imm.dtu.dk

RESUMEN

El análisis arquetípico (AA), propuesto por Cutler y Breiman en [1], estima el envolvente convexo principal de un conjunto de datos. De esta forma, el AA favorece características que constituyen las “esquinas” representativas de los datos, es decir, aspectos o arquetipos distintivos. Mostramos que el AA posee la interpretabilidad del clustering —sin limitarse a asignaciones duras— y la unicidad de la SVD —sin restringirse a representaciones ortogonales—. Para realizar AA a gran escala, derivamos un algoritmo eficiente basado en gradiente proyectado, así como un procedimiento de inicialización inspirado en el método FURTHESTFIRST ampliamente usado en K-means [2]. Demostramos que el modelo de AA es relevante para la extracción de características y reducción de dimensionalidad en una gran variedad de problemas de aprendizaje automático provenientes de visión por computador, neuroimagen, minería de texto y filtrado colaborativo.

1. INTRODUCCIÓN

Los métodos de descomposición se han convertido en una herramienta clave para una amplia gama de análisis de datos masivos: desde el modelado de datos de Internet como matrices término-documento de ocurrencias de palabras, datos bioinformáticos como microarreglos de expresión génica, datos de neuroimagen como actividad neuronal medida en espacio y tiempo, hasta problemas de filtrado colaborativo como el célebre problema de Netflix, por mencionar algunos. Los enfoques convencionales van desde aproximaciones de rango reducido como la descomposición en valores singulares (SVD) y el análisis de componentes principales (PCA) [3], análisis de componentes independientes (ICA) [4], sparse coding (SC) [5] y la factorización de matrices no negativas (NMF) [6], hasta métodos de clustering con asignación dura como K-means y K-medoids.

En general, todos estos métodos pueden entenderse como una representación tipo mezcla lineal o análisis factorial con diversas restricciones. Así, los datos x_{m,n}, donde m = 1, ..., M es el índice de característica y n = 1, ..., N es el índice de muestra, se escriben en términos de variables ocultas s_{k,n} y proyecciones a_{m,k} con k = 1, ..., K:

x_{m,n} = \sum_k a_{m,k} s_{k,n} + e_{m,n}, e_{m,n} ~ N(0, σ^2)

Usualmente con un modelo de ruido gaussiano. La SVD/PCA requiere que A y S sean ortogonales; en ICA se asume independencia estadística para S; en SC se introduce un término de penalización que mide la desviación de la esparsidad en S; y en NMF todas las variables están restringidas a ser no negativas. En K-means, S es una matriz de asignación binaria tal que A = X S^T (S S^T)^{-1} representa los centros euclídeos de cada clúster, mientras que para K-medoids a_k = x_n para algún n, es decir, los centros de clúster deben ser puntos de datos reales.

A pesar de las similitudes de los enfoques mencionados, sus representaciones internas de los datos difieren notablemente y, por lo tanto, también la naturaleza de las interpretaciones que ofrecen. En SVD/PCA, las características representan las direcciones de máxima variación (eigenmaps); en NMF, las características corresponden a partes constituyentes; en SC, los elementos son también átomos o elementos de diccionario; mientras que K-means y K-medoids identifican objetos prototipo más representativos.

Una ventaja de los métodos de clustering es que las características son similares a los datos medidos, lo que facilita su interpretación; sin embargo, las asignaciones binarias reducen la flexibilidad. Además, el clustering suele implicar optimizaciones combinatorias complejas, lo que lleva a una gran variedad de heurísticas. Por otro lado, las aproximaciones de bajo rango como SVD/PCA/NMF ofrecen una gran flexibilidad, pero las características pueden ser más difíciles de interpretar. La invarianza a la rotación de las características extraídas puede provocar falta de unicidad, por ejemplo: X ≈ A S = A Q Q^{-1} S = Ã ■. Además, SVD/PCA/ICA/SC son propensos a efectos de cancelación, donde dos componentes pierden significado porque localmente se vuelven altamente correlacionados, adoptando valores positivos y negativos que casi se anulan, aunque sigan siendo ortogonales globalmente.

En conclusión, los métodos de clustering ofrecen características fáciles de interpretar, pero sacrifican flexibilidad debido a las asignaciones binarias de los objetos de datos. Los enfoques como SVD/PCA/ICA/NMF/SC tienen mayor flexibilidad de modelado y pueden capturar de manera más eficiente, por ejemplo, la varianza, pero esta eficiencia puede dar lugar a representaciones complejas de las que se aprende relativamente poco. El Análisis Arquetípico (AA) propuesto en [1] combina directamente las virtudes del clustering con la flexibilidad de la factorización matricial. En el trabajo original sobre AA [1], el método se demostró útil en el análisis de la contaminación del aire y la forma de la cabeza, y más tarde también para el seguimiento de dinámicas espaciotemporales [7]. Recientemente, el AA se ha utilizado en benchmarking y estudios de mercado para identificar prácticas típicamente extremas [8], así como en el análisis de espectros astronómicos [9] como un enfoque para el problema de extracción de end-members [10]. En este artículo mostramos: (i) que el modelo AA es único; (ii) que puede inicializarse eficientemente mediante el método FURTHESTSUM; y (iii) que puede calcularse eficazmente usando un método simple de gradiente proyectado. Además, mostramos que AA es útil en una gran variedad de dominios, resultando en características fáciles de interpretar que representan bien las dinámicas inherentes en los datos.

2. ANÁLISIS ARQUETÍPICO Y EL ENVOLVENTE CONVEXO PRINCIPAL

El envolvente convexo (también llamado envolvente convexo mínimo) de una matriz de datos X es el conjunto convexo más pequeño que contiene a X. De manera informal, puede describirse como una banda elástica que envuelve a todos los puntos de datos. Aunque el problema de encontrar el envolvente convexo puede resolverse en tiempo lineal (O(N)) [11], el tamaño del conjunto convexo crece dramáticamente con la dimensionalidad de los datos. El tamaño esperado del conjunto convexo para N puntos en posición general en un espacio de dimensión K crece exponencialmente con la dimensión como O(log^{K-1}(N)) [12]. Como resultado, en espacios de alta dimensión, el conjunto convexo “mínimo” que forma el envolvente no proporciona una representación compacta de los datos.

El Análisis Arquetípico considera el envolvente convexo principal, es decir, el envolvente convexo de dimensión (K-1) que mejor representa los datos de acuerdo con alguna medida de distorsión D(·|·). Esto puede formalizarse como la búsqueda de C y S que resuelva: min_{C,S} D(X | X C S) sujeto a: ||c_k||_1 = 1, ||s_n||_1 = 1, C ≥ 0, S ≥ 0. La restricción ||c_k||_1 = 1 junto con C ≥ 0 obliga a que la matriz de características A = X C sea un promedio ponderado (combinación convexa) de las observaciones de datos. Por su parte, la restricción ||s_n||_1 = 1 y S ≥ 0 obliga a que el dato n se aproxime como combinación convexa de los vectores de características X C ∈ R^{M×K}. En lo que sigue consideraremos D(X | X C S) = ||X – X C S||_F^2.

De forma análoga al PCA, los C y S óptimos generarán el envolvente convexo principal (dominante) para los datos X. AA favorece características que representen las “esquinas” de los datos, es decir, aspectos o arquetipos distintivos. Además, el modelo AA puede considerarse naturalmente como un punto intermedio entre las aproximaciones de factorización de bajo rango y los métodos de clustering.

Al igual que K-means y K-medoids, AA es invariante a la escala y traslación de los datos y, como se indicó en [1], el problema AA es no convexo.

2.1. Unicidad del AA

La falta de unicidad en la descomposición matricial es una de las principales motivaciones para el uso de criterios de rotación, como varimax en el análisis factorial, así como para imponer independencia estadística en ICA [4] o esparsidad en SC [5]. A continuación, probamos que el AA es, en general, único salvo por una permutación de los componentes.

Teorema 1.

Supóngase que para todo k existe n tal que c_{n,k} > 0 y c_{n,k'} = 0 para k' ≠ k. Entonces el modelo AA no sufre de ambigüedad rotacional; es decir, si X ≈ X C S = X C Q^{-1} Q S = X C■ S■ y tanto (C,S) como (C■,S■) son soluciones equivalentes, entonces Q es una matriz de permutación.

Demostración.

Dado que S■ = Q S y S ≥ 0 y S■ ≥ 0 son ambas soluciones, se tiene ||s_n||_1 = 1 y ||Q s_n||_1 = 1 para todo n. Para que esto sea cierto, Q debe ser una matriz de Markov (∑_{k'} q_{k,k'} = 1, Q ≥ 0). Dado que C ≥ 0 y C Q^{-1} ≥ 0, y por hipótesis ∀k ∃n : c_{n,k} > 0 y c_{n,k'} = 0 si k' ≠ k, entonces Q^{-1} debe ser no negativa. Si tanto Q como Q^{-1} son no negativas, Q solo puede ser una matriz de escala y permutación; y como la suma por filas de Q debe ser 1, Q solo puede ser una matriz de permutación.

La condición anterior indica que para cada columna de C debe existir una fila donde ese elemento sea el único distinto de cero. Esto se cumple en AA, ya que dos aspectos distintos en posición general no serán combinación convexa de los mismos puntos. Aunque AA es único en general, no hay garantía de identificar la solución óptima debido a mínimos locales.

2.2. Inicialización eficiente del AA mediante FURTHESTSUM

Cutler y Breiman señalan en [1] que una inicialización cuidadosa mejora la velocidad de convergencia y reduce el riesgo de encontrar arquetipos poco significativos. Para K-means, un procedimiento popular es FURTHESTFIRST [2], que inicia con un punto aleatorio y luego elige los siguientes puntos más alejados de los ya seleccionados:

j_nuevo = argmax_i { min_j ||x_i – x_j||, j ∈ C } (2.1)

Para AA proponemos FURTHESTSUM: seleccionar j_nuevo = argmax_i { ∑_{j ∈ C} ||x_i – x_j|| } (2.2). Para mejorar el conjunto C, el primer punto aleatorio se elimina y se selecciona uno adicional en su lugar.

Teorema 2.

Los puntos generados por FURTHESTSUM están garantizados a pertenecer al conjunto convexo mínimo de los puntos no seleccionados.

Demostración.

Por contradicción. Supóngase que existe un punto t fuera del conjunto convexo mínimo, esto es, x_t = X c con ||c||_1 = 1, c_d ≥ 0 y c_t = 0, y que t = argmax_i ∑_{j ∈ C} ||x_i – x_j||. Entonces: ∑_{j ∈ C} ||x_t – x_j|| = ||X c – x_j|| < ∑_d c_d ∑_{j ∈ C} ||x_d – x_j|| ≤ max_d ∑_{j ∈ C} ||x_d – x_j||, siendo la primera desigualdad la triangular. Por lo tanto, existe un punto mejor que t, contradiciendo la optimalidad de t.

Comparación: FURTHESTFIRST distribuye prototipos de forma más uniforme (útil para K-means), mientras que FURTHESTSUM extrae puntos pertenecientes al conjunto convexo de los datos no seleccionados (útil para AA). El costo computacional principal de FURTHESTSUM para identificar T candidatos es O(M N T).

2.3. Gradiente proyectado para AA

Proponemos un procedimiento de gradiente proyectado adaptable a cualquier medida de proximidad; aquí usamos mínimos cuadrados. En [1] se estimó el modelo con mínimos cuadrados no negativos, imponiendo restricciones mediante penalizaciones cuadráticas. También puede resolverse vía programación cuadrática no negativa con restricciones lineales [13]. Sin embargo, el siguiente método funcionó eficientemente en la práctica.

Reformulación invariante a normalización ■1: s■_{k,n} = s_{k,n} / ∑_k s_{k,n}, c■_{n,k} = c_{n,k} / ∑_n c_{n,k}. Con ∂ s■_{k',n} / ∂ s_{k,n} = δ_{k',k} / ∑_k s_{k,n} − s_{k',n} / (∑_k s_{k,n})^2. Las actualizaciones simultáneas son:

Actualización de S: s_{k,n} ← max{ s■_{k,n} + μ_{S■} (g^{S■}_{k,n} − ∑_{k'} g^{S■}_{k',n} s■_{k',n}), 0 }, s■_{k,n} = s_{k,n} / ∑_k s_{k,n}, G^{S■} = C■^T X^T X − C■^T X^T X C■ S■.

Actualización de C: c_{n,k} ← max{ c■_{n,k} + μ_{C■} (g^{C■}_{n,k} − ∑_{n'} g^{C■}_{n',k} c■_{n',k}), 0 }, c■_{n,k} = c_{n,k} / ∑_n c_{n,k}, G^{C■} = X^T X S■^T − X^T X C■ S■ S■^T.

Cada actualización usa búsqueda lineal para μ. En la actualización de S, C■^T X^T X y C■^T X^T X C■ pueden precomputarse (O(K M N)); el gradiente y la evaluación de la función objetivo const. − 2■C■^T X^T X, S■■ + ■C■^T X^T X C■, S■ S■^T■ cuestan O(K^2 N). En la actualización de C, X^T X S■^T y S■ S■^T se precomputan en O(K M N) y O(K^2 N), y el resto es O(K M N).

Si solo se consideran T puntos candidatos para definir arquetipos [13], la complejidad se reduce a O(K M T). En [13] se sugirió identificarlos como outliers proyectando autovectores de la covarianza de X; aquí, FURTHESTSUM es una alternativa eficiente. En nuestra implementación realizamos 10 búsquedas lineales por cada actualización alternada de S y C.

2.4. kernel-AA

Las estimaciones dependen solo de las relaciones por pares (productos internos) K = X^T X. Por tanto, AA se generaliza de forma inmediata a representaciones kernel basadas en relaciones por pares (kernel-AA), interpretándose como la extracción del envolvente convexo principal en un espacio de Hilbert potencialmente infinito (análogo a kernel-K-means y kernel-PCA). Estos análisis quedan fuera del alcance de este trabajo.

3. RESULTADOS

Demostramos la utilidad de AA en cuatro conjuntos de datos procedentes de dominios relevantes de aprendizaje automático.

Visión por Computador.

Comparamos AA con SVD/PCA, NMF y K-means en la base CBCL (361 píxeles × 2429 imágenes) usada en [6]. SVD/PCA extrae características de baja a alta frecuencia espacial; NMF produce una representación basada en partes; K-means obtiene centros de clúster de rostros típicos, mientras que AA extrae prototipos faciales arquetípicos más distintos, exponiendo variabilidad y diversidad. AA explica más variación que K-means pero menos que SVD/PCA y NMF (véase Tabla 1), extrayendo eficientemente aspectos faciales distintivos.

Neuroimagen.

Analizamos un conjunto PET con 40 tiempos × 157,244 vóxeles usando [18F]-Altanserin para medir receptores 5-HT2A. Cada vóxel es mezcla de regiones vasculares, sin unión y de alta unión. Idealmente, AA extrae estos perfiles y cómo cada vóxel es combinación convexa de ellos. En la práctica, AA extrajo tres componentes que corresponden bien a las tres regiones; SVD/PCA, NMF y K-means no lograron perfiles puros y produjeron mezclas.

Minería de Texto.

En el corpus NIPS Bag of Words (1,500 documentos, 12,419 palabras, ~6.4M ocurrencias; normalización IDF), un modelo AA de 10 componentes extrae categorías temáticas distintas. En AA, C indica qué documentos constituyen los aspectos XC; S (no mostrado) indica el grado en que cada documento se asemeja a esos aspectos. Los arquetipos extraídos corresponden bien a tipos de artículos distintos del corpus.

Filtrado Colaborativo.

Analizamos MovieLens mediano (1,000,209 valoraciones de 3,952 películas por 6,040 usuarios) y grande (10,000,054 valoraciones de 10,677 películas por 71,567 usuarios), con valoraciones en {1,2,3,4,5}. AA extrae usuarios idealizados (comportamientos extremos) y relaciona usuarios reales con estos arquetipos. Tratamos películas no calificadas como valores faltantes extendiendo el objetivo de AA:

$$\min_{S,C} \sum_{n,m:\, q_{n,m}=1} \left(x_{n,m} - \left(\sum_k \sum_{m'} x_{n,m'} c_{m',k} \,/\, \sum_{m'} q_{n,m'} c_{m',k} \right) s_{k,m} \right)^2 .$$

Dado que SVD y NMF fueron más propensos a mínimos locales que AA, se inicializaron con la solución de AA. Aunque AA es más restringido que NMF y SVD, logra errores de prueba similares y características mucho más útiles que K-means para predecir valoraciones.

4. DISCUSIÓN

Mostramos que el AA de [1] es útil en diversos problemas de aprendizaje automático. Derivamos un algoritmo sencillo para ajustar AA y el procedimiento FURTHESTSUM para extraer end-members iniciales. La ventaja de AA frente a clustering es que enfatiza aspectos distintivos y mantiene flexibilidad mediante asignaciones suaves. Observamos interpretabilidad mejorada frente a factorizaciones y clustering clásicos. Un problema abierto es determinar el número de componentes (como en SVD/PCA/NMF/SC/K-means), para lo cual pueden usarse criterios de evidencia del modelo o

error de generalización. AA es una herramienta prometedora de aprendizaje no supervisado; su unicidad general lo hace especialmente atractivo para minería de datos.

5. REFERENCIAS

[1] A. Cutler y L. Breiman, “Archetypal analysis,” Technometrics, 36(4):338–347, 1994.

[2] D. S. Hochbaum y D. B. Shmoys, “A best possible heuristic for the k-center problem,” Math. Oper. Res., 10(2):180–184, 1985.

[3] G. H. Golub y C. F. Van Loan, Matrix Computation, 3ª ed., 1996.

[4] P. Comon, “Independent component analysis, a new concept?,” Signal Processing, 36:287–314, 1994.

[5] B. A. Olshausen y D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” Nature, 381:607–609, 1996.

[6] D. D. Lee y H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, 401(6755):788–791, 1999.

[7] E. Stone y A. Cutler, “Introduction to archetypal analysis of spatio-temporal dynamics,” Physica D, 96(1–4):110–131, 1996.

[8] G. C. Porzio, G. Ragozini y D. Vistocco, “On the use of archetypes as benchmarks,” Appl. Stoch. Models Bus. Ind., 24(5):419–437, 2008.

[9] B. H. P. Chan, D. A. Mitchell y L. E. Cram, “Archetypal analysis of galaxy spectra,” Mon. Not. R. Astron. Soc., 338:790, 2003.

[10] P. Perez, R. Plaza, J. Plaza, A. Martinez, “A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data,” IEEE TGRS, 42(3):650–663, 2004.

[11] D. McCallum y D. Avis, “A linear algorithm for finding the convex hull of a simple polygon,” Information Processing Letters, 9:201–206, 1979.

[12] R. A. Dwyer, “On the convex hull of random points in a polytope,” Journal of Applied Probability, 25(4):688–699, 1988.

[13] C. Bauckhage y C. Thurau, “Making archetypal analysis practical,” Proc. 31st DAGM Symposium on Pattern Recognition, 272–281, 2009.