Reconocimiento de Patrones y Aprendizaje Automatizado

Análisis de conjuntos de datos usando agrupamiento por clustering

M. CONCHA VÁZQUEZ
L. HERNÁNDEZ CANO
M. X. LEZAMA HERNÁNDEZ
E. E. VÁZQUEZ SALCEDO

Facultad de Ciencias, UNAM

Facultad de Ciencias
Universidad Nacional Autónoma de México
http://www.fciencias.unam.mx/

Título del proyecto:
Agrupación de semillas de trigo con base algunas de sus características geométricas.

Tema del proyecto:
Aprendizaje no supervisado; Clustering.

Periodo:
Abril, 2018.

Número de Grupo:
7085

Participantes:
Concha Vázquez Miguel.
Hernández Cano Leonardo.
Lezama Hernández María Ximena.
Vázquez Salcedo Eduardo Eder.

Supervisores:
Profesor Gustavo De la Cruz Martínez.
Rafael Robles Ríos.
Estefania Prieto Larios.

Fecha en que se completó el proyecto:
30 de abril del 2018.

Resumen:

Aplicamos técnicas de aprendizaje automatizado no supervisado para el aglomeramiento de semillas de trigo basándonos en características geométricas y cuantitativas de éstas. En particular empleamos *cluserting* jerárquico y K-MEANS en una implementación que directamente ocupa funciones provistas por el lenguaje de programación de propósito estadístico R, haciendo un análisis comparativo de las diferentes técnicas usadas.

Índice general

Capítulo 1

Planteamiento

1.1. Introducción

Desde el inicio de la agricultura, la base de gran parte de la sociedad ha sido el trigo. Junto con el maíz y el arroz. el trigo es una de los granos más ampliamente producidos globalmente[2]. Lo más destacado de los cereales es que sus frutos maduros son no perecederos y pueden ser almacenados para consumirse gradualmente o mantenidos como semilla. Actualmente el el análisis de semillas es una actividad muy importante en el proceso de producción de granos.

En los últimos años se han llevado a cabo análisis muy diversos en cuanto a las caracterśiticas físicas que exhiben las semillas de trigo y en gran medida los núcleos de éstos con tal de poder argrupar a diferentes tipos de semillas con base en estas observaciones como se muestra en [1][3][5]. El prosito es claro: luego de haber cultivado el trigo, la preocupación principal en la industria es la clasificación de acuerdo a su tamaño, calidad, variedad, entre otras cuestiones. Una amplia varidad de técnicas que miden estos parámetros y analizan las semillas de trigo consumen mucho tiempo y son propensos al error humano. Para ésto se han desarrollado sistemas automatizados que clasifiquen y agrupen las semillas más rápidamente y sin estar sujetos al factor humano. En este trabajo se desarrolla uno de estos sistemas con tal de analizar la factibilidad de estas propuestas y originar nuevas ideas.

1.2. Definición del problema

Las características de una semilla están ligadas a la especie de planta a la que pertenece la semilla. Dada una especie de planta y un cierto número de semillas de la planta se espera observar patrones que relacionen sus características.

Si tuviéramos un montón de semillas esperaríamos poder agruparlas de forma que las semillas que pertenecen a un mismo grupo tienen similitud entre sus características y por tanto pertenecen a la misma especie. Buscamos entonces inferir una clasificación de un conjunto de semillas a partir de patrones estadísticamente significativos entre sus atributos.

Para un tratamiento formal del problema definamos algunos conceptos.

Definición 1 (Semilla) *Una semilla de dimensión n es un vector* $[a_1, ..., a_n] \in R^n$. *Decimos que las entradas de la semilla son sus* ***atributos***.

En nuestro caso particular trabajaremos con el caso $n = 7$, y los n atributos representarán el área, perímetro, longitud del núcleo, ancho del núcleo, coeficiente de asimetría y longitud del surco del núcleo, de la semilla.

Definición 2 (Clasificación de un conjunto de semillas) *Dado un conjunto S de semilas de dimensión n, una clasificación de S es una partición indexada $C_1, ..., C_m$ de S. Decimos que la clasificación de $s \in S$ es i si $s \in C_i$.*

Nos interesa encontrar un proceso que para cualquier conjunto de semillas nos devuelva una clasificación que agrupe las semillas en conjuntos ajenos con relación estadísticamente significativa entre los atributos de las semillas de un mismo conjunto.

Definición 3 (Clasificador de semillas) *Un clasificador de semillas de dimensión n es una función que toma un conjunto de semillas S y devuelve una clasificación C de S tal que las relaciones entre los atributos de las semillas de cada conjunto de C son estadísticamente significativas.*

Notemos que la definición 3 no es matemáticamente precisa, pues la noción de *relación estadísticamente significativa* no está bien definida. Es parte del problema encontrar una medida de similitud entre semillas que resulte apropiada de acuerdo a alguna otra medida.

Es importante que el proceso realice inferencias sobre correlación entre los valores de un atributo de forma automática, pues las semillas no están etiquetadas.

Debido al gran número de ejemplares necesarios para obtener una clasificación estadisticamente significativa es deseable diseñar un proceso que infiera de alguna forma las características que debe de tener un clasificador de semillas, pues no resulta viable que esta tarea sea realizada por un ser humano en tiempo razonable.

Definición 4 (Clasificación no supervisada) *Cuando se infiere una clasificación a partir de los atributos de los ejemplares sin una clasificación a priori se realiza una clasificación no supervisada.*

De acuerdo con esta definición el proceso que realizaremos para obtener el clasificador de semillas es un ejemplo de aprendizaje no supervisado.

Capítulo 2

Metodología

Clustering puede considerarse el problema de aprendizaje no supervisado más importante; entonces, como cualquier otro problema de este tipo, se trata de encontrar una estructura en una colección de datos sin etiqueta. Una definición amplia de clustering podría ser ël proceso de organizar objetos en grupos cuyos miembros son similares de alguna manera". Un clúster es, por lo tanto, una colección de objetos que son *"similares"* entre ellos y son *"diferentes"* a los objetos que pertenecen a otros grupos.

2.1. Descripción del clustering

Clustering es una técnica de minería de datos (*data mining*) dentro de la disciplina de Inteligencia Artificial que identifica de forma automática agrupaciones o *clústeres* de elementos de acuerdo a una medida de similitud entre ellos. El objetivo fundamental de las técnicas de clustering consiste en identificar grupos o clústeres de elementos tal que:

- La similitud media entre elementos del mismo clúster sea alta. **Similitud intra-clúster alta**.
- La similitud media entre elementos de distintos clústeres sea baja. **Similitud inter-clúster baja**.

2.1.1. Agrupamiento mediante *k-means*

K-means clustering es un tipo de aprendizaje no supervisado, que se utiliza cuando se tiene datos no etiquetados (es decir, datos sin categorías o grupos definidos). El objetivo de este algoritmo es encontrar grupos en los datos, con el número de grupos representados por la variable **K**. El algoritmo funciona iterativamente para asignar cada punto de datos a uno de *los grupos K* en función de las características que se proporcionan. Los puntos de datos se agrupan según la similitud de características. Los resultados del algoritmo de agrupamiento K-means son:

- Los centroides de los clústeres K, que se pueden usar para etiquetar nuevos datos.
- Etiquetas para los datos de entrenamiento (cada punto de datos se asigna a un solo grupo)

En lugar de definir grupos antes de mirar los datos, la agrupación le permite buscar y analizar los grupos que se han formado orgánicamente.

Cada centroide de un clúster es una colección de valores de características que definen los grupos resultantes. Examinando los pesos de las características del centroide se puede interpretar cualitativamente el tipo de grupo que representa cada clúster.

El algoritmo de agrupamiento de K-means usa refinamiento iterativo para producir un resultado final. Las entradas de algoritmo son el número de clústeres K y el conjunto de datos. El conjunto de datos es una colección de características para cada punto de datos. Los algoritmos comienzan con estimaciones iniciales para los K centroides, que pueden generarse aleatoriamente o seleccionarse aleatoriamente del conjunto de datos. El algoritmo luego itera entre dos pasos:

1. Paso de asignación de datos:

   Cada centroide define uno de los clusters. En este paso, cada punto de datos se asigna a su centroide más cercano, en función de la distancia euclidiana al cuadrado. Más formalmente, si c_i es la colección de centroides en el conjunto C, entonces cada punto de datos x se asigna a un clúster basado en:

   $$\text{argmin } dist(c_i, x)^2 \text{ tal que } c_i \in C$$

   donde $dist()$ es la distancia euclidiana estándar y permite que el conjunto de asignaciones de puntos de datos para cada i-ésimo centroide de grupo sea S_i.

2. Paso de actualización del centroide:

   En este paso, los centroides se vuelven a calcular. Esto se hace tomando la media de todos los puntos de datos asignados al clúster de ese centroide.

   $$c_i = \frac{1}{|S_i|} \sum x_i \in S_i$$

   El algoritmo itera entre los pasos uno y dos hasta que se cumple un criterio de detención (es decir, ningún punto de datos cambia los clústeres, la suma de las distancias se reduce al mínimo o se alcanza un número máximo de iteraciones).

   Se garantiza que este algoritmo converge a un resultado. El resultado puede ser un óptimo local (es decir, no necesariamente el mejor resultado posible), lo que significa que evaluar más de una ejecución del algoritmo con centroides iniciales aleatorizados puede dar un mejor resultado.

El algoritmo encuentra los clústeres y las etiquetas de los conjuntos de datos para una K particular elegida previamente. Para encontrar la cantidad de conglomerados en los datos, necesitamos ejecutar el algoritmo de agrupamiento K-means para un rango de valores K y comparar los resultados . En general, no existe un método para determinar el valor exacto de K, pero se puede obtener una estimación precisa usando las siguientes técnicas.

Una de las métricas que se usa comúnmente para comparar los resultados en diferentes valores de K es la distancia media entre los puntos de datos y su centroide de grupo. Dado que aumentar el número de clústeres siempre *reducirá* la distancia a los puntos de datos, al aumentar K siempre disminuirá esta métrica, hasta el extremo de llegar a cero cuando K es igual que la cantidad de puntos de datos.

Por lo tanto, esta métrica no se puede usar como el único objetivo. En cambio, si se traza la distancia media al centroide como una función de K y el *"punto de inflexión"*, donde la velocidad de disminución cambia bruscamente, se puede usar para determinar aproximadamente K.

2.1.2. Agrupamiento mediante *clusters jerárquicos*

La técnica de clustering jerárquico construye un dendograma o árbol que representa las relaciones de similitud entre los distintos elementos. La exploración de todos los posibles árboles es un problema NP. Por lo tanto, suelen seguirse algoritmos aproximados guiados por determinadas heurísticas.

La agrupación jerárquica, como su nombre lo sugiere, funciona en base un algoritmo que construye una jerarquía de clusters. Este algoritmo comienza con todos los puntos de datos asignados a un clúster propio. Luego, dos clústeres más cercanos se fusionan en el mismo clúster. Al final, este algoritmo termina cuando solo queda un solo grupo y los resultados del agrupamiento jerárquico se pueden mostrar con dendrograma.

En un dendrograma, en la parte inferior, comenzamos con n puntos de datos, cada uno asignado a clusters separados. Dos clusters más cercanos se fusionan hasta que tenemos un solo grupo en la parte superior. La altura en el dendrograma en el que se fusionan dos clústeres representa la distancia entre dos clústeres en el espacio de datos.

La decisión del número de clusters que pueden representar mejor a los diferentes grupos se pueden elegir al observar el dendrograma. La mejor elección del número de clusters es el número de líneas verticales en el dendrograma cortadas por una línea horizontal que puede atravesar la distancia máxima verticalmente sin intersectar un grupo.

El algoritmo se puede dar con un enfoque de abajo hacia arriba, pero también es posible seguir un enfoque descendente que comience con todos los puntos de datos asignados en el mismo clúster y realizar divisiones recursivas hasta que a cada punto de datos se le asigne un clúster separado.

La decisión de fusionar dos clusters se toma sobre la base de la cercanía de estos clusters. Hay múltiples métricas para decidir la cercanía de dos clusters, así como técnicas de similitud para el aglomeramiento jerárquico que discutimos en la implementación[4][6]. Dos de las métricas más comunes para hacer el cálculo de cercanía son la distancia euclidiana,

$$\|a - b\|_2 = \sqrt{\sum a_i - b_i}$$

y la distancia de Manhattan

$$\|a - b\|_1 = \sum |a_i - b_i|$$

.

2.2. Los clusters como herramienta para el análisis de conjuntos de datos

De acuerdo con la descripción anterior del agrupamiento por *clustering*, esta forma de inferir una clasificación a partir del conjunto de datos nos permite realizar una exploración de los datos sin conocimiento previo sobre ellos, pues su resultado es una partición de elementos cuyos elementos tienen una representación vectorial cercana.

En nuestro caso particular las semillas son desde un inicio vectores, y por la justificación dada en la sección introductoria esperamos que las particiones que los procesos de agrupamiento nos dan tengan relación con la clasificación de las semillas por especie de planta a la que pertenecen.

De esta forma utilizamos al agrupamiento por *clustering* como una forma de explorar el conjunto de datos sin indicar al proceso la clasificación que esperamos obtener, sin embargo en este caso podemos comparar las particiones resultantes con la clasificación que tenemos de antemano.

2.3. Ventajas del análisis mediante agrupamiento

- Si se analiza un conjunto de datos sobre el que no se tiene ninguna etiqueta, este análisis nos permitiría encontrar patrones estadísticos en el conjunto de datos.
- No hace falta tener una muestra clasificada *a priori* por la naturaleza no supervisada del procedimiento de agrupación.
- Es una técnica sencilla de utilizar, pues a diferencia de otras no hace ninguna suposición sobre los datos.
- Sólo hace falta representar a los ejemplares como vectores para poder aplicar esta técnica.

2.4. Desventajas del análisis mediante agrupamiento

- La clasificación resultante trata únicamente de cercanía entre la representación vectorial de los ejemplares. La interpretación de los resultados se puede volver complicada, pues su valor semántico no es necesariamente evidente ni significativo para el objetivo del análisis.
- La falta de suposiciones sobre el conjunto de datos limita de cierta forma el tratamiento formal de los resultados, esto hace difícil la interpretación de la clasificación.
- La representación vectorial de los ejemplares en este caso fue trivial, sin embargo esta situación no se da en todos los problemas. Esto limita la aplicabilidad de la técnica.
- Si la cercanía vectorial entre las representaciones de dos ejemplares no tiene análogo semántico para los ejemplares originales, entonces la técnica no tiene sentido.
- La complejidad del proceso para agrupar por *k-means* es $O(n^{dk+1}logn)$, donde k es el número de grupos esperados y d la dimensión del espacio de vectores de los ejemplares. Al igual que en el caso del clasificador Bayesiano-no ingenuo, esta cota limita su aplicabilidad.

Notemos que la última desventaja listada es importante: si la cercanía de la representación de dos ejemplares no implica una cercanía semántica de algún tipo de los ejemplares originales, no hay ninguna razón para suponer que los patrones estadísticos de las presentaciones tendrán significado para los datos originales. Esto pone en evidencia la importancia de la representación vectorial de los ejemplares.

Capítulo 3

Descripción de la propuesta e implementación

3.1. Implementación del sistema

Previo a analizar propiamente el código y las estructuras que fueron necesarias definir para llevar a cabo el *clustering* de los datos con base en las observaciones cuantitativas de las semillas, describamos adecuadamente a los ejemplares u observaciones sobre las cuales trabajamos para estudiar su estructura.

3.1.1. Obtención de ejemplares

Decidimos desde un inicio contar con un conjunto de observaciones no tan grande como en proyectos previos dado que la complejidad del algoritmo de *k-means* es $\mathcal{O}(n^{dk+1}logn)$ en donde k es dijo y es el número de entidades a particionar y d (también fijo) la dimensión del espacio en donde se hace el agrupamiento, además de ser $NP-duro$ en caso de que no estén dijos estos parámetros. Por otro lado, la complejidad de los algoritmos de agrupamiento jerárquico es del orden $\mathcal{O}(n^3)$-tiempo con n el número de observaciones consideradas.

Luego de haber motivado el proyecto, nos referimos a un conjunto de observaciones (*dataset*) de Data.World, que a si vez fue transformado a formato *CSV* por Jonathan Ortiz a partir del repositorio de aprendizaje de máquina (UCI, *Machine Learning Repository*)[7].

La razón de la elección de este conjunto de datos se debió a la cantidad de observaciones recopiladas (210), que teníamos las etqieutas de clase para poder efectuar después un análisis comparativo de los agrupamientos que obtendríamos (3 etiquetas) y la uniformidad o preprocesamiento a partir del cual se midieron las propiedades físicas de las semillas de trigo en la que ahondamos en la sección del conjunto de datos.

Descripción del conjunto de datos

En el conjunto de datos se incluyen obervaciones para tres tipos de semillas de trigo, *Kama*, *Rosa* y *Canadiense* que fueron seleccionadas de forma aleatoria para el expermiento, pero procurando que hubiera un total de setente (70) observaciones para cada tipo de semilla. Como consecuencia, en total se tienen doscientes diez (210) observaciones de semillas de trigo en donde los atributos o propiedades geométricas medidas fueron el área, perímetro,

longitud del núcleo, ancho del núcleo, coeficiente de asimetría y longitud del surco del núcleo, de cada semilla. Además —como se mencionó antes— cada observación viene etiquetada con el tipo de semilla del que se trata[1].

Las mediciones fueron llevadas a cabo por el Instituto de Agrofífica de la Academia Polaca de Ciencias en Lublin usando un método de visualización de alta calidad del núcleo interno de las semillas con ayuda de una ténica suava de rayos-X no destructiva de la semilla que se considera más económica y que da mejores resultados que la microscopía u otras técnicas de láser. Cada imágen se grabó en placas de KODAK para rayos-X de 13x18 centímetros

3.1.2. Implementación en R

Con tal de poder tener un buen análisis con conjunto con el código de la implementación, aprovechamos la característica de los cuadernos de `R` con soporte para *markdown* proporcionado por `R Studio`. De esta forma es posible ver una explicación detallada de cada fase de la implementación, su argumento e ir ejecutando por trozos el código del *script* para ir obteniendo la salida. Más aún, como separamos en dos archivos —uno para *k-means* y el otro para el *agrupamiento jerárquico*— es posible convertir el cuaderno a formato `HTML` o incluso `PDF` usando la herramienta de *knitting* provista por `R Studio`.

Por esta razón, las implementaciones concretas y las decisiones particulares del código las especificamos en estos *PDFs* que fueron generados. Es posible cargar los archivos con extensión `.Rmd` directamente en `R Studio` para ver el cuaderno a partir del cual fueron generados, además de poder elegir la opción de replicar como nosotros el expermento (incluso fijando para ésta tarea la semilla de aleatoriedad) o bien elegir como tal nuevos parámetros para llevar a cabo nuevos experimentos y al finalizar elegir la opción de generar un nuevo `PDF` con los resultados del nuevo experimento. Nótese que son necesarios los paquetes `rmarkdown`, `knitr` y `readr` para la ejecución del programa, así como tener el conjunto de datos en la dirección especificada en la documentación.

Consultar el archivo de documentación del proyecto en R para más información.

- *k-means*.
- *hierachical-clustering*.

3.1.3. Resultados

A grandes rasgos, pudimos ver que es posible determinar relativamente con facilidad el número de *clusters* que es conveniente usar para los algoritmos de *k-means* y de *clustering jerárquico* con base en una gráfica que compare el número de clusters contra la inercia inter-clusters (*square sum withinss*) (también conocida como suma de cuadrados inter-grupos) para determinar el punto en que el incremento ya no es significativo y podemos cortar en dado punto. De esta forma podemos asegurar la mayor heterogeneidad de los *clusters* formados sin sacrificar el aspecto de que incrementar el número de grupos aumenta la complejidad y puede degenerar en tener un cluster para cada observación en el pero de los casos. Sin embargo, como se discutió en los cuadernos de `R`, esto depende en gran parte de cada problema y el objetivo del estudio.

También nos fue posible extrapolar de *k-means* y el *clustering jerárquico* que es complicado diferenciar al primer tipo de semillas *kama* de las *canadienses*, mientas que el desjuste entre comparaciones entre la especie de semillas de trigo *rosa* con los otros dos fue menor.

[1] Por esta razón y al ser la ténica de *clustering* un tipo de aprendizaje no serupervisado, tuvimos que crear una copia de los datos del conjunto y quitarle la columna correspondiente a estas etiquetas, aunque posteriormente nos sirvieron para hacer un análisis comparativo.

Para más información acerca de los resultados obtenidos, consultar la documentación del proyecto en R provista por los cuadernos antes mencionados[2].

[2]En caso de que existan problemas al ejecutar los trozos de código de los cuadernos, será fundamental que los paquetes y bibliotecas sugeridas estén instaladas en el sistema; esto puede efectuarse con el instalador de paquetes de `R Studio`.

Capítulo 4

Conclusiones

El análisis realizado permitió explorar las técnicas utilizadas y apreciar los motivos detrás de algunas de sus limitaciones. En particular observamos que el agrupamiento por *clustering* puede tener buenos resultados como fue en el caso del conjunto de datos de semillas.

4.1. Análisis mediante agrupamiento por *clustering*

Estas técnicas resultan útiles para realizar una exploración inicial de un conjunto de datos no etiquetado, pues al no hacer suposiciones del conjunto se puede aplicar a una gran variedad de situaciones siempre que se tenga una representación vectorial apropiada de los ejemplares.

La interpretación de los resultados puede resultar difícil, sin embargo es posible que devuelva un resultado útil. Desafortunadamente no es necesariamente evidente encontrar una representación vectorial de los ejemplares que resulte apropiada.

En resumen, la aplicación directa de las técnicas exploradas resulta útil únicamente cuando la representación vectorial de los ejemplares es tal que la cercanía de la representación de dos ejemplares implica similitud de los ejemplares originales, y aún cuando se cumple esa situación puede resultar difícil encontrar el criterio de semejanza en los ejemplares originales.

En el conjunto de datos analizado en este proyecto las preocupaciones expuestas están mitigadas por el hecho de que los ejemplares ya eran vectores y la similitud entre los objetos que representaban estaba bien modelada por la distancia entre los vectores. Por estas razones el análisis ofreció los resultados esperados.

4.2. Conjunto de datos

Observamos que las propiedades de las semillas son indicadores más o menos apropiados de las semillas a las que representan. Esta conclusión se justifica observando que contábamos con una clasificación del conjunto de datos explorados, lo cuál nos permitió comparar los resultados obtenidos con los resultados esperados.

Esto parece indicar que las técnicas exploradas tienen su mayor potencial de utilidad en aquellos conjuntos de datos cuya representación vectorial sea natural, como fue en el caso de las semillas.

Bibliografía

[1] Fourth International Conference on Computational e Information Sciences. *Wheat Cultivar Classifications Based on Tabu Searchiand Fuzzy C-means Clustering Algorithm*.

[2] International Journal of Computer Applications. *Autonomous Wheat Seed Type Classifier System*.

[3] International Conference on Computer Science y Electronics Engineering. *Design and Realization of Grain Seed Quality Testing System Based on Particle Image Processing Technology*.

[4] Universidad de Granada. *Métodos Jerárquicos de Análisis de Cluster*.

[5] International Conference on Image Processing e its Applications. *A PC based colour image processing system for wheat grain grading*.

[6] Science Foundation Ireland. *Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm*.

[7] Institute of Mathematics y Konstantyn Computer Science The John Paul II Catholic University of Lublin. *Seeds Dataset, Measurements of geometrical properties of kernels belonging to three different varieties of wheat*.