Información Tecnológica
Vol. 30(2), 135-142 (2019)
http://dx.doi.org/10.4067/S0718-07642019000200135

Innovación en Minería de Datos para el Tratamiento de Imágenes: Agrupamiento K-media para Conjuntos de Datos de Forma Alargada y su Aplicación en la Agroindustria

Trung T. Pham[1]*, Gustavo A. Lobos[2] y Cristian L. Vidal-Silva[3]
(1) Centro de Investigación en Tecnología de Información, Facultad de Economía y Negocios, Universidad de Taca, Av. Lircay S/N, Talca-Chile. (e-mail: tpham@utalca.cl)
(2) Centro de Mejoramiento y Fenómica Vegetal, Facultad de Ciencias Agrarias, Universidad de Talca, Av. Lircay S/N, Talca – Chile. (e-mail: globosp@utalca.cl)
(3) Ingeniería Civil Informática, Escuela de Ingeniería, Campus Rodelillo, Universidad Viña del Mar, Agua Santa 7055, Viña del Mar-Chile. (e-mail: cristian.vidal@uvm.cl)

* Autor a quien debe ser dirigida la correspondencia

Recibido Jun. 14, 2018; Aceptado Ago. 17, 2018; Versión final Ago. 25, 2018, Publicado Abr. 2019

Resumen

Este trabajo presenta un innovador método de agrupación K-media modificado basado en la teoría de conjunto junto con su aplicación en el ámbito de procesamiento de imágenes agroindustrial. K-media tradicional permite la agrupación de conjuntos en subconjuntos mediante la definición de centros según la fórmula de distancia. Cuando los datos se concentran en formas sin un sentido hiper-esférico, esta herramienta permite que el centro del conjunto, con un único punto, se convierta en un subconjunto de muchos puntos. En este artículo se presenta una modificación de la fórmula de distancia que permite dar mayor flexibilidad para el estudio de casos en agricultura. Mediante ejemplos numéricos, la funcionalidad y aplicabilidad del método modificado de agrupación K-media es evaluada en imágenes infrarrojas provenientes de ensayos de déficit hídrico en trigo.

Palabras clave: agrupación K-media; teoría de conjuntos; función de distancia; monitoreo de estrés

Innovation in Data Mining for the Image Processing: K-means Clustering for Data Sets of Elongated Forms and its Application in the Agroindustry

Abstract

This paper presents an innovative modified method of K-means clustering based on the set theory together with its application in the processing images of the agroindustry field. Traditional K-means permits the clustering of sets in subsets by means of defining their center according to the distance formula. When the data is concentrated in forms without a hyper-spherical sense, this tool allows the center of the set, with a single point, to become a subset of many points. In this article we present a modification of the distance formula that allows giving more flexibility for the study of cases in agriculture. Using numerical examples, the functionality and applicability of the modified method of K-means grouping is evaluated in infrared images from water deficit tests in wheat.

Keywords: K-means clustering; set theory; distance function; environmental stress monitoring

INTRODUCCIÓN

El agrupamiento o clustering (Aggarwal, 2014; Celebi, 2015; Han et al., 2011) es una técnica de dividir un conjunto de datos en muchos subconjuntos, de manera que cada uno sólo contenga datos con cierto grado de similitud. Para datos numéricos, esta semejanza está dada por la distancia entre los mismos en el especio de datos: los más similares son aquellos con la distancia mínima entre sí. Como un punto de referencia, el centro de cada grupo de datos es aquel punto para el cual, de manera implícita, los datos que están cerca de este centro también lo están entre ellos. En este sentido, el método de agrupación K-media (Wu, 2011; He et al., 2013; Chen et al., 1998) se desarrolló para identificar grupos de datos que se pueden analizar para propósitos más profundos tales como clasificación, extracción de información relevante, y descubrimiento de información nueva, entre otros (Han et al., 2011; Zabalza et al., 2015; Tsang et al., 2015). Si bien el método K-media puede implementarse fácilmente en una aplicación o software informático, el principal desafío radica en que la función de distancia requiere de la concentración uniforme de los datos alrededor del centro.

En este sentido, esta metodología podría ser útil para el análisis de imágenes infrarrojas en agricultura. Sin embargo, las regiones de interés poseen formas alargadas que hacen difícil una identificación efectiva de cada una de ellas. No obstante lo anterior, el método K-media no es del todo inútil porque parte de cada grupo todavía se identifica alrededor de un punto central, y cuando los datos se concentran alrededor de líneas paraleles es posible, aunque por suerte que ellas son simétricas, que los grupos se identifiquen correctamente. Debido a que la función de distancia entre dos puntos en verdad es un caso especial de la función de distancia entre dos conjuntos (Kaplan, 2001; Desgupta, 2014), se hipotetiza que sería factible modificar el método K-media, de manera de incrementar su flexibilidad en el estudio de imágenes infrarrojas para identificar regiones de interés (follaje) y descartar aquellas que no son de utilidad (suelo, aire, etc.). De mismo modo, el análisis de otros tipos de datos, por ejemplo datos económicos, que busca datos de forma alargada para identificar modelo de regresión lineal puede beneficiarse con esta modificación del método K-media.

Este artículo presenta como innovación la extensión del método K-media de agrupación de un caso especial al concepto general, con la función de distancia definida para medir distancia entre dos conjuntos de datos en lugar de dos puntos de datos. Cuando estos conjuntos son líneas, la función de distancia se convierte en una formula especifica de forma cerrada. Con esta fórmula, la agrupación K-media se extiende para subconjuntos con centros de forma de una línea (en lugar de forma de un punto). Además, el procedimiento de actualizar el centro de un subconjunto después de una iteración se extiende para el caso en que el centro es una línea en lugar de un punto. En esta extensión, la actualización se realiza a través del método numérico de mínimos cuadrados (Cheney y Light, 2009). Este trabajo es una extensión de trabajo (Pham y Lobos, 2017).

PROPUESTA Y RESULTADOS

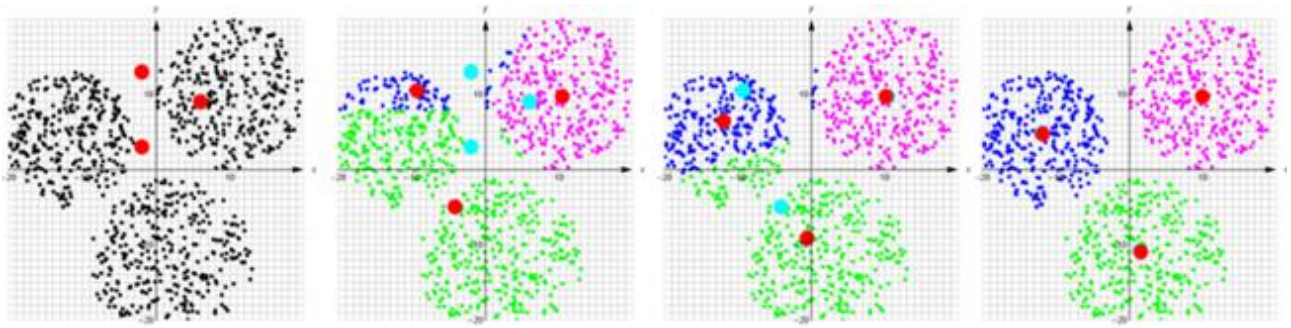
La agrupación K-media es normalmente efectivo cuando los datos se ubican en concentración alrededores de un punto central, formando un subconjunto con forma esférica (o hiper-esférica) en un espacio de datos. Este requerimiento implícito es debido al uso de la función de distancia D(p1, p2) entre dos puntos de datos $p_1 = (r_{1,1}, r_{1,2}, r_{1,3}, \ldots, r_{1,K})^T$ and $p_2 = (r_{2,1}, r_{2,2}, r_{2,3}, \ldots, r_{2,K})^T$, respectivamente:

$$D(p_1, p_2) = \sqrt{\sum_{k=1}^{K} (r_{1,k} - r_{2,k})^2} \ . \quad (1)$$

El concepto de agrupación K-media es sencillo: un número de subconjuntos se determina según la iniciación de los centros de estos subconjuntos. En relación con la distancia más corta entre cada punto y su centro, cada dato es asignado a un subconjunto. Una vez realizado el paso previo, los centros de estos subconjuntos se actualizan, y el proceso se repite hasta que no existen nuevas actualizaciones de los centros de los subconjuntos. La figura 1 presenta ejemplos del proceso de agrupación K-media de un subconjunto de datos distribuidos circularmente en el espacio de dos dimensiones. En este proceso, cada gráfica representa los resultados después de una iteración en el proceso (puntos celestes). El centro de cada subconjunto se muestra en el color rojo, y los puntos de cada subconjunto se muestran en el color asignado (verde, azul, y rosa, respectivamente). En este ejemplo, el centro de cada subconjunto se logró después de tres iteraciones. Cuando los datos se concentran de una manera diferente a lo anteriormente expuesto (figura 2), la agrupación K-media es menos eficaz.

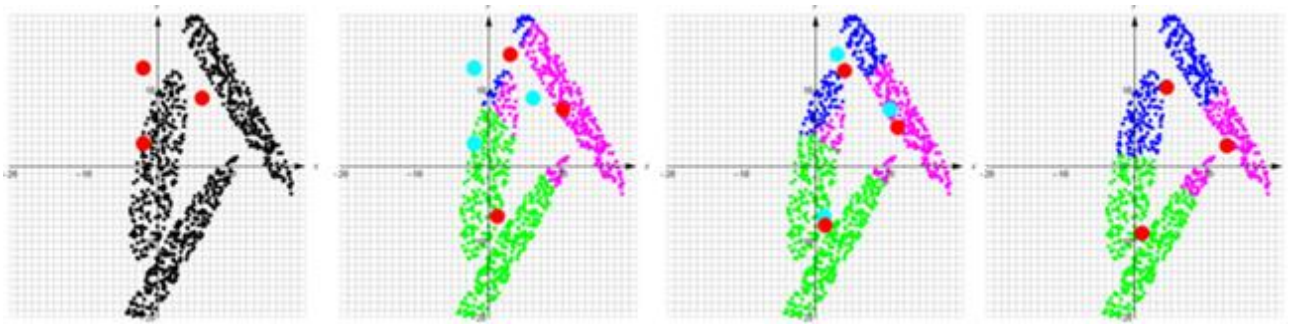
En este ejemplo, el proceso termina después de cinco iteraciones, y cada subconjunto final contiene puntos de dos subconjuntos reales (diferentes colores) implicando que durante el proceso no fue posible una completa separación de los puntos. En este ejemplo, el problema se origina en la forma alargada de los

subgrupos, además de la proximidad de sus bordes (lo que complica la individualización y separación de los mismos). Así, surge la necesidad de una modificación que permita trabajar con datos que poseen este tipo de distribución. En lugar de desarrollar una técnica nueva, es quizás una solución más eficiente, la aplicación de mejoras sobre la actual metodología, de manera de alcanzar una mayor flexibilidad.



(a) Punto de inicio (no separación) (b) Iteración 1 (3 grupos) (c) Iteración 2 (tres grupos) (d) Iteración 3 (tres grupos)

Fig. 1: Ejemplo de proceso iterativo de agrupación K-media.



(a) Punto de inicio (no separación) (b) Iteración 1 (3 grupos) (c) Iteración 2 (tres grupos) (d) Iteración 3 (tres grupos)

Fig. 2: Ejemplo de proceso de agrupación K-media con subconjuntos de forma.

En este sentido, el centro de un conjunto (o subconjunto) se define como un conjunto de puntos. Cuando este conjunto contiene un sólo punto, se corresponde con los casos tradicionales en donde se aplica K-media. La función de distancia $D(\cdot,\cdot)$ entre dos conjuntos S_1 y S_2 es definida de acuerdo a la teoría de conjuntos como sigue:

$$D(S_1, S_2) = \min_{p_1 \in S_1, p_2 \in S_2} D(p_1, p_2), \tag{2}$$

donde la función de distancia entre dos puntos ya se definió en (1). Así, cuando ambos conjuntos, S_1 y S_2, contienen un sólo punto, la función (2) se simplifica a (1). En el contexto de la extensión de la agrupación K-media, cuando S_1 es un punto $p = (p_1, p_2)$, y S_2 es una línea de forma $S_2 = L = \{ (r_1, r_2) \mid \alpha r_1 + \beta r_2 = \gamma \}$, la función de distancia en (2) se convierte en un problema de optimización constreñido:

$$\min_{r_1, r_2} \ (p_1 - r_1)^2 + (p_2 - r_2)^2, \tag{3a}$$

$$\alpha r_1 + \beta r_2 = \gamma, \tag{3b}$$

que, por la sustitución de (3b) en (3a), se convierte en un problema de optimización no constreñido, con una solución única debido de la naturaleza de la función convexa resultante, y esta solución se obtiene en forma cerrada:

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix}^* = \frac{1}{\alpha^2 + \beta^2} \begin{bmatrix} \beta^2 & -\alpha\beta \\ -\alpha\beta & \alpha^2 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} + \frac{\gamma}{\alpha^2 + \beta^2} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \tag{4}$$

Sustituyendo (4) en (3a), y la función de objetivo en (3a) se optimiza y representa la solución del problema (2):

$$D(p, L) = \frac{(-\gamma + \alpha p_1 + \beta p_2)^2}{\alpha^2 + \beta^2}, \tag{5}$$

Como la distancia entre el punto $p = (p_1, p_2)$, y el conjunto $L = \{ (r_1, r_2) \mid \alpha r_1 + \beta r_2 = \gamma \}$. Esta función de distancia se utilizará en el primer paso del algoritmo de la agrupación K-media, para posteriormente actualizar el centro de cada subconjunto. En este caso, el centro de una subconjunto es una línea de forma $L = \{ (r_1, r_2) \mid \alpha r_1 + \beta r_2 = \gamma \}$. Dado los puntos en un subconjunto $S_k = \{ (p_{1,1}, p_{1,2}), (p_{2,1}, p_{2,2}), \ldots, (p_{N,1}, p_{N,2}) \}$, el centro de este subconjunto, en la forma de una línea $L = \{ (r_1, r_2) \mid \alpha r_1 + \beta r_2 = \gamma \}$, se actualiza en el cálculo de las constantes α, β y γ determinando la línea L:

$$\min_{\alpha,\beta} \sum_{n=1}^{N} (\gamma - \alpha p_{n,1} - \beta p_{n,2})^2 \tag{6}$$

Debido de la naturaleza convexa en (6), el problema tiene una solución única cuando el constante α se normaliza a 1, implicando que la línea no es horizontal:

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix}^* = \begin{bmatrix} \sum_{n=1}^{N} p_{n,2}^2 & -\sum_{n=1}^{N} p_{n,2} \\ -\sum_{n=1}^{N} p_{n,2} & \sum_{n=1}^{N} 1 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{n=1}^{N} p_{n,1} p_{n,2} \\ \sum_{n=1}^{N} p_{n,1} \end{bmatrix}. \tag{7a}$$

Cuando la línea es horizontal, (7a) es numéricamente singular, significando que el constante α es cero y no se debe normaliza a 1, y en este caso el constante β se normaliza a 1, implicando que la línea no es vertical:

$$\begin{bmatrix} \alpha \\ \gamma \end{bmatrix}^* = \begin{bmatrix} \sum_{n=1}^{N} p_{n,1}^2 & -\sum_{n=1}^{N} p_{n,1} \\ -\sum_{n=1}^{N} p_{n,1} & \sum_{n=1}^{N} 1 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{n=1}^{N} p_{n,1} p_{n,2} \\ \sum_{n=1}^{N} p_{n,2} \end{bmatrix}. \tag{7b}$$

Combinando (5) con (7a), o como en caso (7a) es numéricamente singular, con (7b), el algoritmo de la agrupación K-media se modifica para el caso en que un centro de un subconjunto siendo una línea. En el presente planteamiento, podría darse el caso en que líneas se crucen entre sí (Fig. 3b-d), en la intersección de dos líneas generadas donde los puntos de datos se pueden asignar incorrectamente a un clúster. Producto de lo anterior, a veces es mejor reemplazar el uso de líneas por el de segmentos para la identificación de los centros de cada subconjunto de datos.

Cuando un segmento se utiliza como un centro de un conjunto de datos, éste se representa por dos puntos de extremo. Por lo tanto, durante la ejecución de (5), se debe considerar (4) para determinar si el resultado está entre dos puntos de extremo, y después de calcular (7), los dos puntos de extremo de la línea se deben actualizar. Para determinar si el resultado r^* de (4) está entre dos puntos de extremo r_A y r_B, calcular δ:

$$r^* = r_A + \delta(r_B - r_A), \tag{8}$$

y si $0 \le \delta \le 1$, se puede concluir que r^* está entre r_A y r_B. Cuando r^* está afuera de r_A y r_B, la función de distancia (5) se debe modificar:

$$D(p, L) = \begin{cases} \dfrac{(-\gamma + \alpha p_1 + \beta p_2)^2}{\alpha^2 + \beta^2}, & \text{si } 0 \le \delta \le 1, \\ \min(D(r^*, r_A), D(r^*, r_B)), & \text{de otro caso.} \end{cases} \tag{9}$$

La figura 3 ilustra gráficamente la ecuación (9) en el cálculo de la función de distancia entre un punto y un segmento de línea. Cuando la proyección de un punto de dato en la línea es dentro del segmento (el caso de $0 \le \delta \le 1$), la distancia correcta es la distancia entre el punto de dato y el punto de su proyección en la línea. Cuando la proyección de un punto de dato en la línea es afuera del segmento (el caso de $\delta < 0$ o de $\delta > 1$), la distancia correcta es el mínimo de la distancia entre el punto de dato a los puntos de extremo del segmento.

Para actualizar los dos puntos de extremo r_A y r_B, primero las líneas se actualizan utilizando (7). Después, las proyecciones (4) de los puntos de cada subconjunto se comparan en un par que resultan en la más larga distancia entre sí. Estos puntos se asignan como puntos de extremo del segmento central. La figura 4 muestra algunos resultados del algoritmo modificado de la agrupación K-media con la función de distancia en (5) entre un punto y una línea, y la actualización del centro de un subconjunto en (7) para el centro siendo una línea. Como se aprecia, las líneas son capaces de generar tres subconjuntos que son, evidentemente, más homogéneos que los formados en la figura 2. Un pequeño problema mostrado en este ejemplo es en el área alrededor de una intersección entre dos líneas centrales: en teoría cuando un punto es proyectado en dicha intersección, se puede asignar a ambos subconjuntos indicados por estas líneas centrales. En la implementación, se asigna arbitrariamente a la primera línea encontrada en un arreglo. Por esta razón, se sugiere utilizar segmentos en lugar de líneas como centros de subconjuntos. La figura 5 muestra algunos resultados del algoritmo modificado de la agrupación K-media con la función de distancia en (9) entre un punto y un segmento. Estos resultados muestran que el problema en la intersección de dos líneas se mejora.

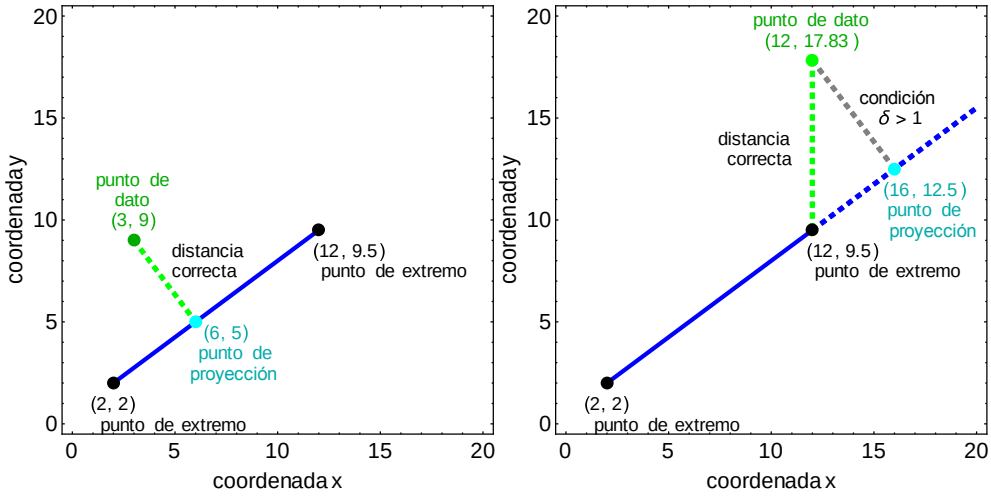
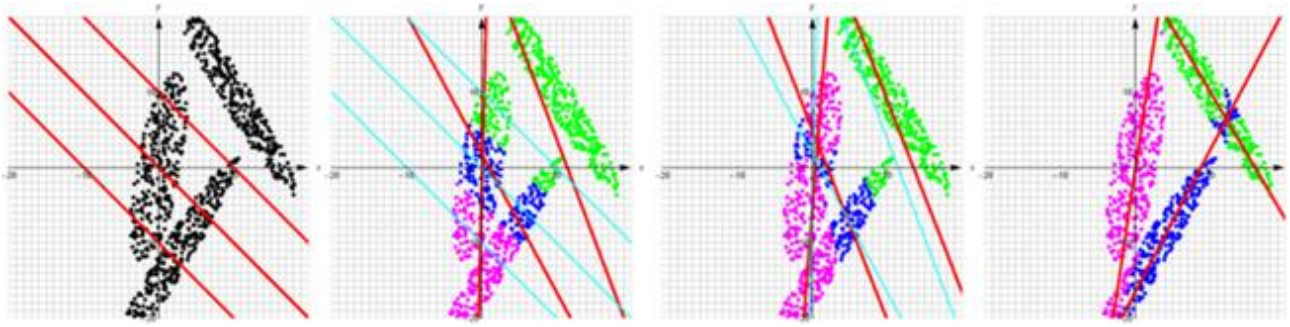
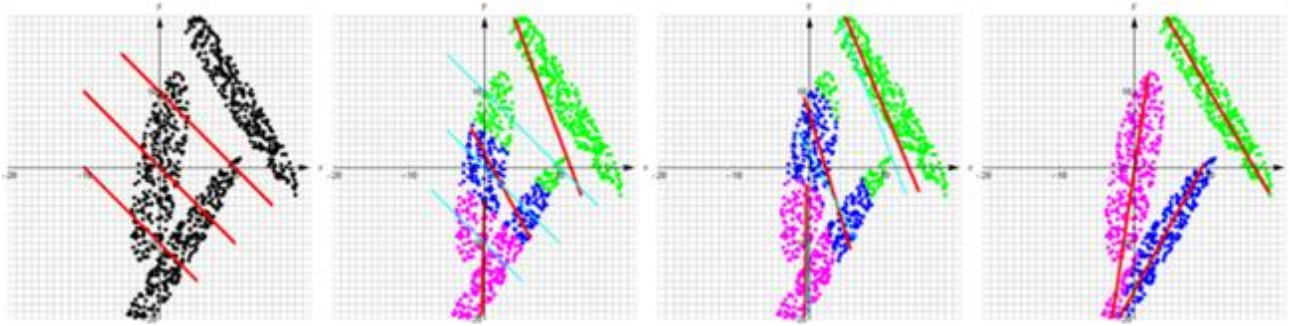


Fig. 3: Hay dos casos en la fórmula de distancia entre un punto y un segmento de línea: (a) cuando la proyección de dato es dentro del segmento, y (b) cuando la proyección de dato es afuera del segmento.



(a) Punto de inicio (no separación) (b) Iteración 1 (3 grupos) (c) Iteración 2 (tres grupos) (d) Iteración 28 (tres grupos)

Fig. 4: Ejemplo de proceso modificado de agrupación K-media.



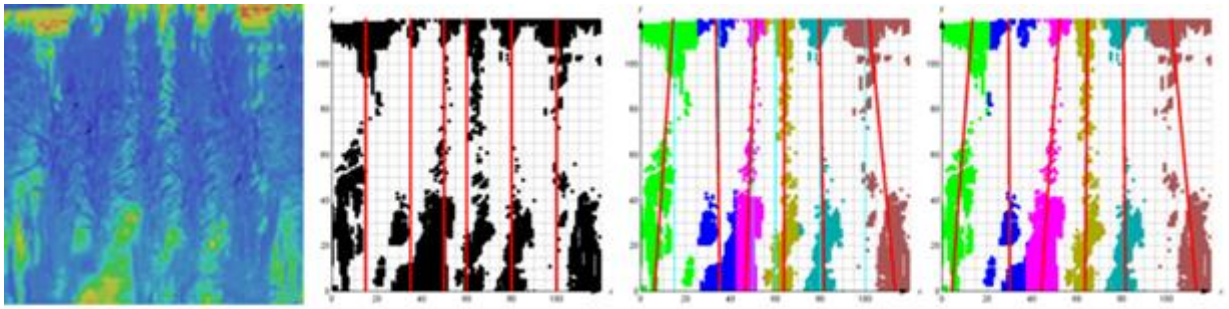
(a) Punto de inicio (no separación) (b) Iteración 1 (3 grupos) (c) Iteración 2 (tres grupos) (d) Iteración 20 (tres grupos)

Fig. 5: Ejemplo de proceso de agrupacion K-media modificado.

Innovación y Aplicación Agrícola con Imágenes Infrarrojas

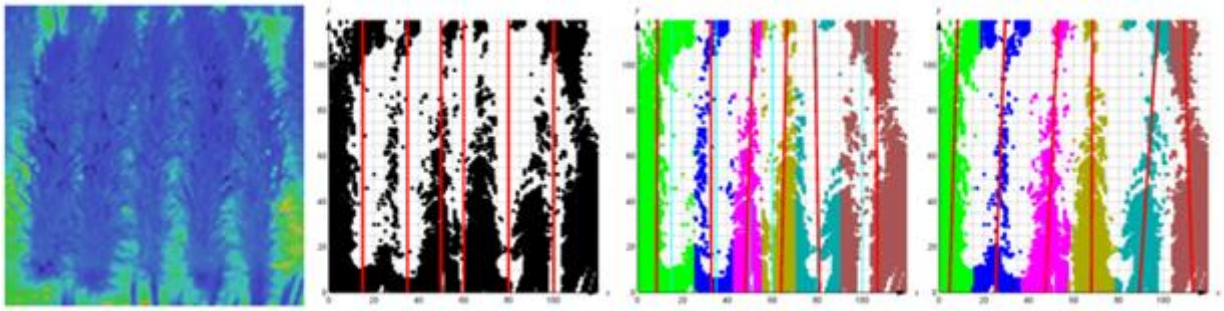
En esta sección, se presenta una aplicación del sector agroindustrial utilizando el método modificado de agrupación K-media presentado en el apartado previo. En esta aplicación, imágenes infrarrojas se tomaron en el campo para estudiar el efecto del déficit hídrico sobre la temperatura del follaje de plantas de trigo. En este ensayo de campo, cada parcela (2 x 1 m) fue fotografiada con una cámara de imagen infrarroja (FLIR i-40, Flir System Inc., OR, USA).

En cada imagen (parcela) se aprecian cinco filas que corresponden a las plantas de trigo, pero al mismo tiempo y producto de las diferencias entre parcelas, en algunas imágenes también se observan plantas de las parcelas vecinas, lo que dificulta la automatización del análisis termal. A lo anterior se suma la necesidad de remover el suelo de cada una de las imágenes. Además, después de remover las partes no relevantes, se necesita dividir el bloque cuadrado que contiene cinco filas (hileras de trigo) de plantas para estudiarlas individuamente. Las figuras 6(a), 7(a), 8(a) y 9(a) son imágenes típicamente encontradas en este estudio, en donde se encuentran parcelas en las cuales es fácil identificar las cinco hileras, pero en otras esta labor se hace más difícil.



(a) Imagen original (b) Punto de inicio (no separación) (c) Iteración 1 (seis grupos) (d) Iteración 5 (seis grupos)

Fig. 6: Ejemplo 1 de identificacion de franjas límites en imágenes infrarrojas.



(a) Imagen original (b) Punto de inicio (no separación) (c) Iteración 1 (seis grupos) (d) Iteración 5 (seis grupos)

Fig. 7: Ejemplo 2 de identificacion de franjas límites en imágenes infrarrojas.

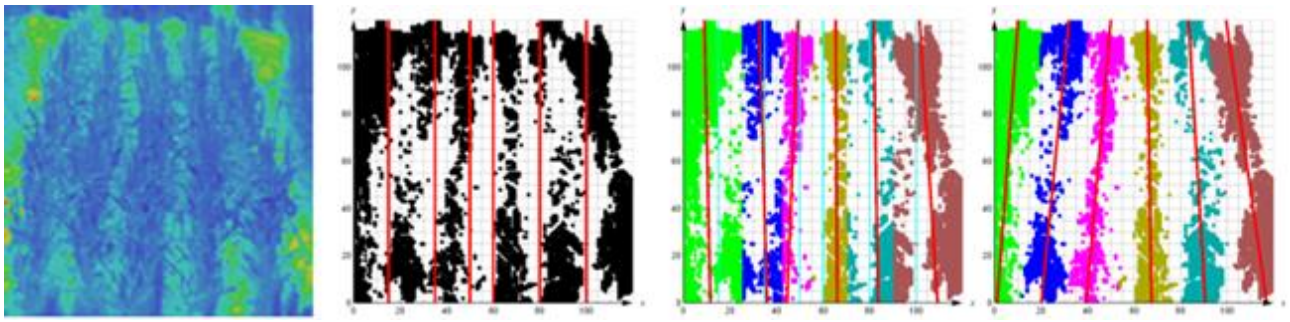
En el primer paso de preprocesamiento de datos, temperaturas altas (asociando con la tierra) y temperaturas bajas (asociando con las plantas) se identifican y separan a través del análisis de histograma de una imagen. Después, los pixeles conteniendo temperaturas altas se modelan como grupos de datos en forma alargada similar a los ejemplos en la previa sección. Utilizando el método modificado de agrupación K-media, los pixeles conteniendo temperaturas altas se dividen en subconjuntos alargados. Después, estos subconjuntos alargados se utilizan como guías de límite para dividir los pixeles de temperaturas bajas en representaciones de filas individuales de plantas. Las estrechas franjas de tierra a veces se cubren por las plantas, haciendo más difícil el manejo de las imágenes. Por esta razón, el método modificado de agrupación K-media en la previa sección es particularmente útil en el sentido que cada grupo de datos se identifica con una línea representando su centro. *Esta línea se utiliza* cuando no hay pixel disponible para usar como guía de límite continua: la línea de centro se proyecta en pixeles y estos pixeles se llenan en el espacio blanco para crear una guía de límite continua.

DISCUSIÓN

Se ha modificado el algoritmo de agrupación K-media para identificar subconjunto con su centro en la forma alargada. Este algoritmo se ha implementado, y los resultados de casos controlados se presentan para

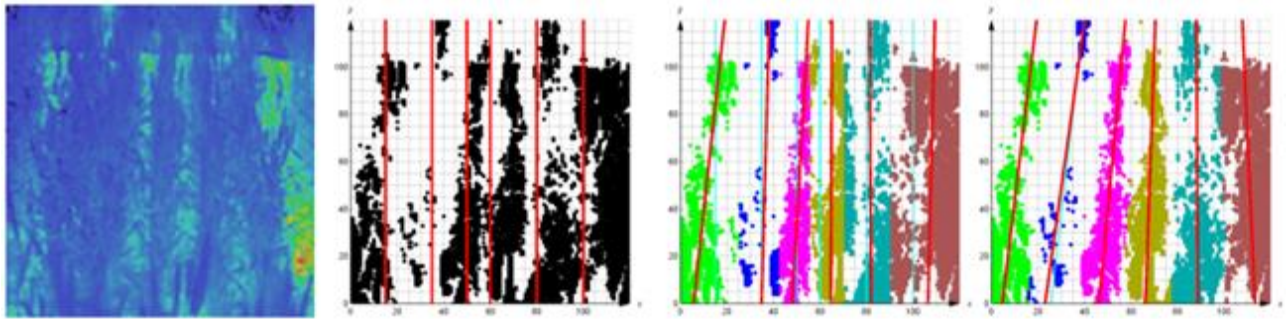
confirmar la funcionalidad de dicho algoritmo. En aplicaciones agrícolas, este algoritmo modificado permitiría identificar las regiones de interés, tal como se describe en este trabajo. Mientras que esta orientación permite analizar de mejor manera las imágenes termales, también genera una pregunta filosófica importante: como el propósito de la minería de datos es descubrir información, ¿es posible desarrollar una técnica que automáticamente permita descubrir la forma de cada subconjunto y de la misma forma decidir qué modificación sobre la función es necesaria para determinar el centro de un subconjunto? Justamente, tal y como presentan las figuras 10 y 11, los resultados obtenidos permitirían a los investigadores orientarse a un estudio más detallado de las temperaturas foliares.

Es importante observar que los resultados de las figuras 8, 9, 10, 11 se obtienen solo con el método K-media modificado porque el método K-media existente entregará resultados de forma circula debido al uso un punto para representar el centro de un clúster que resultará en los clústeres de forma circula. Otros métodos de agrupamiento como método jerárquico no puede separar clústeres que no tienen brechas entre sí, o método basado en modelo también no puede identificar clústeres que no tienen distribución normal como mostradas en dichas figuras. Es también importante observar que otros tipos de datos con clústeres de forma elongada representable por un modelo regresión lineal comunmente encontrados en estudios economicos se pueden analizar con este método K-media modificado.



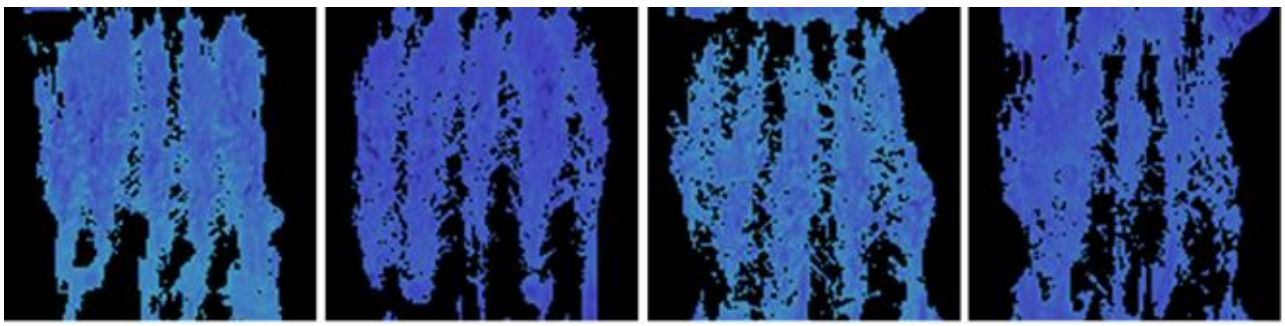
(a) Imagen original (b) Punto de inicio (no separación) (c) Iteración 1 (seis grupos) (d) Iteración 5 (seis grupos)

Fig. 8. Ejemplo 1 de identificacion de franjas límites en imágenes infrarrojas con medio de agrupación K-media modificado.



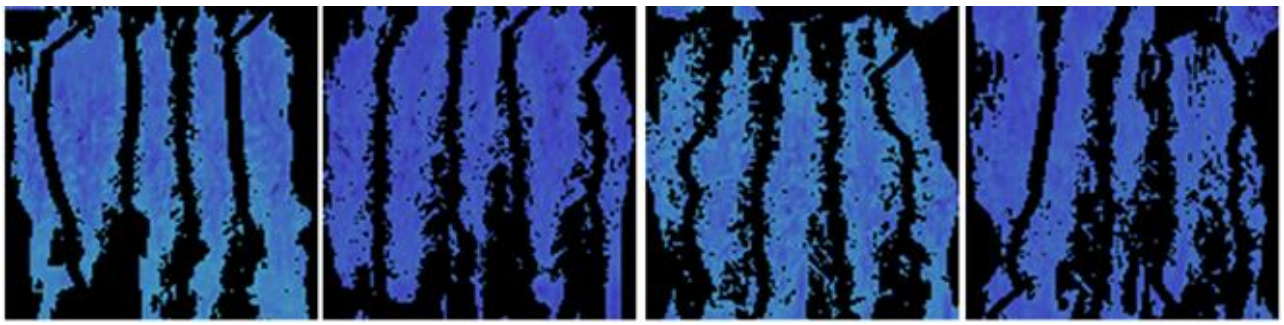
(a) Imagen original (b) Punto de inicio (no separación) (c) Iteración 1 (seis grupos) (d) Iteración 5 (seis grupos)

Fig. 9. Ejemplo 2 de identificacion de franjas límites en imágenes infrarrojas con medio de agrupación K-media modificado.



(a) Cinco filas de figura 6(a) (b) Cinco filas de figura 7(a) (c) Cinco filas de figura 8(a) (d) Cinco filas de figura 9(a)

Fig. 10: Ejemplo 1 de identificacion de franjas límites en imágenes infrarrojas de figuras 6, 7, 8 y 9.



(a) Cinco filas de figura 6(a) (b) Cinco filas de figura 7(a) (c) Cinco filas de figura 8(a) (d) Cinco filas de figura 9(a)

Fig. 11: Ejemplo 2 de identificacion de franjas límites en imágenes infrarrojas de figuras 6, 7, 8 y 9 con método modificado de agrupación K-media.

CONCLUSIONES

Este trabajo ha presentado la aplicabilidad de técnicas de minería de datos para el procesamiento de imágenes en un contexto de la agroindustria en Chile las que usualmente presentan una forma alargada. Justamente, este trabajo presentó la propuesta de aplicar el algoritmo K-media para el agrupamiento o clustering de datos sobre imágenes con forma alargada, con la que se obtienen resultados prometedores, lo cual representa un avance teórico y práctico en el área. Los códigos de programación, los datos, y los resultados en este trabajo son disponibles en el sitio Web del Centro de Investigación en Tecnología de Información de la Universidad de Talca (http://iie.utalca.cl/CITI/project/thermalImages).

AGRADECIMIENTOS

Parte de este trabajo estuve patrocinada por la Comisión Nacional de Investigación Científica y Tecnología (CONICYT) mediante el programa Fondo de Fomento al Desarrollo Científico y Tecnológico, proyecto FONDEF IDEA 14I10106.

REFERENCIAS

Aggarwal, C. C., Data classification: algorithms and applications, Chapman & Hall/CRC, ISBN-13: 978-1466586758, Boca Raton, FL (2014)

Chen, C.W., J. Luo y K.J. Parker, Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications, doi: 10.1109/83.730379, IEEE Transactions on Image Processing, 7 (12) 1673-1683 (1998)

Cheney, W. y W. Light, A course in approximation theory, American Mathematical Society, ISBN-13: 978-0821847985, Providence, RI, USA (2009)

Celebi, M.E., Partitional clustering algorithms, Springer, ISBN-13: 978-3319092584, New York, USA (2015)

Dasgupta, A., Set theory: with an introduction to real point sets, Springer, ISBN-13: 978-1461488538, New York, USA (2014)

Han, J., M. Kamber y J. Pei, Data mining: concepts and techniques, Morgan Kaufmann, ISBN-13: 978-0123814791 (2011)

He, J., C. Kim y C. Jay Kuo, Interactive segmentation techniques: algorithms and performance evaluation, Springer, ISBN-13: 978-9814451598, New York, USA (2013)

Kaplansky, I., Set theory and metric spaces, AMS Chelsea Publishing, ISBN-13: 978-0821826942, New York, USA (2001)

Pham, T. y G. Lobos, Agrupación K-Media de Conjuntos de Datos de Forma Alargada, International Conference on Business Administration and Economy (ICBAE-2017), Universidad de Talca, Chile (2017)

Tang, Y., C. Ten, C. Wang y G. Parker, Extraction of energy information from analog meters using image processing, doi: 10.1109/TSG.2015.2388586, IEEE transactions on smart grid, 6 (4) 2032-2040 (2015)

Wu, J., Advances in K-means clustering: a data mining thinking, Springer-Verlag, ASIN: B010DPZNA8, Berlin, Alemania (2012)

Zabalza, J., J. Ren, J. Zheng, J. Han, H. Zhao, S. Li y S. Marshall, Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging, doi: 10.1109/TGRS.2015.2398468, IEEE transactions on geoscience and remote sensing, 53 (8), 4418-4433 (2015)