ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/311457198

Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva

Conference Paper · November 2016

CITATIONS
0

READS
33

5 authors, including:

Ana Cristina Umaquinga
Universidad Técnica del Norte
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Diego Peluffo
Universidad Técnica del Norte
95 PUBLICATIONS **98** CITATIONS
SEE PROFILE

Juan Carlos Alvarado Pérez
Universidad de Salamanca
19 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Andres Javier Anaya Isaza
South Colombian University
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Semantic Interpretation of Brain Signals View project

A Low Cost Portable Prototype for Support of Bio-Feedback Therapies in Psychophysiological Disorders View project

All content following this page was uploaded by Ana Cristina Umaquinga on 07 December 2016.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva

Visual Big Data analysis proposal using interactive dimensionality reduction

A. C. Umaquinga-Criollo
D. H. Peluffo-Ordóñez
Universidad Técnica del Norte
Ibarra, Ecuador
acumaquinga@utn.edu.ec
dhpeluffo@utn.edu.ec

M. V. Cabrera-Álvarez
J. C. Alvarado-Pérez
A. J. Anaya-Isaza
Corporación Universitaria Autónoma de Nariño
Pasto, Colombia
Universidad Surcolombiana
Neiva, Colombia
milton.cabrera@aunar.edu.co
jcalvarado@usal.es
andres.anaya@usco.edu.co

***Resumen*—**En la actualidad se puede evidenciar un crecimiento exponencial del volumen de datos, dando lugar al área emergente denominada Big Data. Paralelamente a este crecimiento, ha aumentado la demanda de herramientas, técnicas y dispositivos para almacenar, transmitir y procesar datos de alta dimensión. La mayoría de metodologías existentes para procesar datos de alta dimensión producen resultados abstractos y no envuelven al usuario en la elección o sintonización las técnicas de análisis. En este trabajo proponemos una metodología de análisis visual de Big Data con principios de interactividad y controlabilidad de forma que usuarios (incluso aquellos no expertos) puedan seleccionar intuitivamente un método de reducción de dimensión para generar representaciones inteligibles para el ser humano.

Palabras Clave*—Big Data, reducción de dimensión, análisis visual*

***Abstract*—** Today, the volumen of available data is experiencing an exponential growing, introducing an emergent are so-called Big Data. Along with the data growing, the demand of tools, techniques and devices to store, transmit and process high-dimensional data (HD) is increased. Most available methodologies to process HD output abstract outcomes and user is not involved in the selection or parameter tuning processes of data analysis techniques. In this work, we propose a visual analysis methodology following principles of interactivity and controlability. Doing so, users (even non-expert ones) can inuitively select a dimensionality reduction method to generate inteligible representations for human beings.

Palabras Clave*—Big Data, reducción de dimensión, análisis visual*

I. Introducción

El crecimiento del volumen de datos de diferente tipo (estructurados, no estructurados, semiestructurados) es exponencial y actualmente en términos de almacenamiento alcanza el orden de petabytes, y exabytes. Dichos datos son generados por diferentes fuentes, entre ellas: Los seres humanos, la comunicación máquina a máquina (también denominada como M2M), los grandes datos transaccionales, la información biométrica [1], [2], entre otros. El gran volumen de información se debe a los avances electrónicos e informáticos, como sensores, satélites, bandas magnéticas, GPS, tecnologías web, cloud computing, y redes sociales [3], [4].

Uno de los desafíos del manejo de información que presenta el mercado es analizar, descubrir y entender más allá de lo que sus procesos y herramientas tradicionales reportan sobre su información [1]. En efecto, si la información no puede ser fácilmente interpretada, se genera un mayor consumo de recursos tecnológicos, económicos, tiempo, y talento humano (presencia requerida de expertos en análisis de datos).

Las técnicas comunes de tratamiento de datos no permiten recuperar la información oculta en su totalidad o no tienen la capacidad para tratarlos, en consecuencia la visualización de datos en muchos casos se vuelve imprescindible, en especial, en las etapas de análisis en donde se realizan las hipótesis

significativas sobre los datos [5] [4], de forma que los usuarios (no necesariamente expertos) puedan obtener representaciones visuales que permitan analizar de forma intuitiva los resultados [6], [4].

Para lograr descubrir el conocimiento inmerso en Big Data (grandes volúmenes de datos), la visualización de datos/información (denominado VI/ IV/ Info Vis/ Data Vis) pretende representar los datos de forma inteligible. Una forma de lograr esto, es a través de técnicas de reducción de dimensión (RD), que le permite transformar los datos en representaciones visuales de objetos en 1, 2 ó 3 dimensiones [7] desde el punto de vista de la percepción human [8] y podría representar una mejora substancial en el costo computacional.

Uno de los mayores problemas que enfrenta la representación visual es la alta dimensionalidad o dimensión, es decir, un número significativamente grande de variables o atributos que caracterizan a un objeto. Además, las herramientas de VI /IV, en su mayoría, implican etapas de pre procesamiento, uso de métodos de minería de datos como rol importante [9], [7], [10], [11], post procesamiento y/o la visualización. Sin embargo, no todas las herramientas integran todas las etapas mencionadas, terminando en resultados abstractos de la información. Asimismo, las herramientas que integran todas las etapas no tienen especial énfasis en la visualización, por lo que los resultados, a pesar de que involucran un análisis visual, también tienden a ser abstractos [12], [13] o ambiguos, y tan sólo algunas pocas pueden usarse sin conocimiento a priori acerca de los datos [14], [4].

En el presente trabajo, se presenta una metodología de visualización interactiva y eficaz de datos, usando un modelo matemático-geométrico de combinación de técnicas kernel no supervisadas de reducción de dimensión que presente un buen compromiso entre el desempeño en la representación de los datos y el costo computacional.

Esta metodología de visualización interactiva de datos combina diferentes métodos de RD no supervisados y representados en matrices kernel, y permite realizar la mezcla de métodos de forma interactiva, a través de una combinación lineal de las correspondientes matrices kernel cuyos coeficientes se relacionan con las coordenadas geométricas de los puntos interiores de una determinada figura geométrica de tal forma que sea de uso fácil e intuitivo inclusive para un usuario no experto, ya que le permite seleccionar un método específico o combinarlos de acuerdo a sus necesidades .

El resto del documento se organiza como se explica a continuación: En la Sección II, se realiza una breve revisión de los métodos de reducción de dimensión existentes. En la Sección III, se presenta la explicación de la metodología de visualización propuesta. En la Sección IV, aspectos de discusión. Por último, en la Sección V se expone las principales conclusiones de la investigación realizada.

II. BREVE REVISIÓN DE LOS MÉTODOS DE REDUCCIÓN DE DIMENSIÓN EXISTENTES

Entre los métodos clásicos de RD, se encuentra el análisis de componentes principales -*principal component analysis* (PCA) y *classical multidimensional scaling* (CMDS), los cuales se basan en criterios de conservación de la varianza y la distancia, respectivamente [15]. Recientemente, los métodos de RD se enfocan en criterios orientados a la preservación de la topología de los datos. Normalmente, dicha topología se representa mediante una matriz de similitud o afinidad que representa el grado de relación o conexión entre los puntos coordenados (coordenadas cartesianas que representan los datos). Desde un punto de vista de teoría de grafos, los datos pueden representarse a través de un grafo ponderado (grafo con un valor de peso por cada adyacencia o arista) y no dirigido, en el cual los nodos representan los puntos coordenados, y la matriz de similitud o afinidad contiene los pesos de cada arista. Los métodos pioneros en incluir similitudes son *Laplacian eigenmaps* (LE) [16] y *locally linear embedding* (LLE) [17], los cuales son de tipo espectral, es decir que usan la información de los valores vectores y vectores propios.
Por otra parte, dado que la matriz de similitud normalizada puede interpretarse como distribuciones de probabilidad, han surgido otros enfoques basados en divergencias, tales como *stochastic neighbour embedding* (SNE) [18], y sus variantes y mejoras, tales como t-SNE que usa una distribución *t-Student* y JSE que usa la divergencia de Jensen-Shanon. [19] [20].

III. METODOLOGIA DE VISUALIZACIÓN PROPUESTA

En esta propuesta se presenta la integración sinérgica de dos áreas: Reducción de dimensión y visualización de información. Específicamente, se propone un nuevo sistema de visualización basado en reducción de dimensión, siguiendo las reglas de la percepción humana, en las cuales se tiene en cuenta el color, la intensidad, la luminosidad, el sombreado, el brillo, el contraste, la textura, la forma, la orientación, el movimiento, la estereoscopia, entre otros conceptos, para proponer un diseño visualmente significante para la cognición humana. Este nuevo método de visualización interactiva de datos consiste en la combinación de diferentes métodos de reducción de dimensión no supervisados y representados en matrices kernel.

Dicha combinación se basa en un modelo matemático-geométrico que permita realizar la mezcla de métodos de forma interactiva, a través de una combinación lineal de las correspondientes matrices kernel cuyos coeficientes se relacionan con las coordenadas geométricas de los puntos interiores de una determinada figura geométrica. Así, un usuario -incluso, no experto- podría fácil e intuitivamente seleccionar un método en específico o realizar una combinación de métodos que satisfaga sus necesidades por medio de la exploración de una figura geométrica y de la selección de puntos de la superficie de la misma. En Figura 1 se muestra gráficamente la aplicación de un posible modelo matemático-geométrico con un enfoque basado en polígonos, en donde, en general los métodos son representados por un conjunto de funciones $\{f_1, \ldots, f_M\}$, donde *M* es el número de funciones.

Una manera de mezclar dos funciones es la deformación continua de una función en otra, usando principios básicos de homotopía[21]. Un modelo simple de homotopía es

$h(f_1, f_2, \lambda) = \lambda f_1 + (1-\lambda) f_2$, donde λ es un parámetro de homotopía, que en términos de una interfaz serviría de barra deslizante. Gráficamente, este modelo podría representarse como una línea de longitud 1 trazada entre dos puntos que representan las funciones, como se aprecia en la Figura 1(a). Este modelo podría extenderse naturalmente a más de dos métodos de forma que tres funciones se representarían con un triángulo (Figura 2(b)) cuatro funciones con un rombo (Figura 1(c)), y así sucesivamente. Para efectos de visualización de datos a través de métodos de reducción de dimensión, los términos a combinar serían matrices kernel correspondientes a los métodos. Por tanto, la matriz $\widehat{K}$ resultante de la mezcla de un conjunto de M matrices kernel $\{K^{(1)}, \ldots, K^{(M)}\}$ podría escribirse como $\widehat{K} = \sum_{m=1}^{N} \alpha_m K^{(m)}$, donde α_m es la ponderación correspondiente al método *m*. Los coeficientes de ponderación deberán estar asociados con las coordenadas geométricas de los puntos al interior de la superficie del polígono. En la Figura 2 se muestra un ejemplo del modelo con 4 métodos y sus correspondientes parámetros geométricos. El parámetro de homotopía λ, los niveles de profundidad al interior de la superficie $\{\mu_1, \ldots, \mu_n\}$ y el parámetro de resolución de profundidad ε determinan los coeficientes α_m.

Como se mencionó previamente, el método KPCA permite obtener espacios de baja dimensión a través de cualquier método de RD, siempre y cuando éste último pueda representarse adecuadamente en una matriz kernel. No obstante, la ejecución del algoritmo KPCA y el cálculo de las matrices kernel pueden significar un costo computacional elevado de acuerdo con la complejidad de los datos, y por tanto podría afectarse el propósito de lograr una interactividad síncrona con el usuario. Además, la representación resultante de los datos debe ser altamente controlable de forma que se ajuste lo mejor posible a los criterios y necesidades del usuario.

Dicho esto, el método propuesto de visualización de datos debe alcanzar un buen compromiso entre desempeño en la representación de los datos y costo computacional.

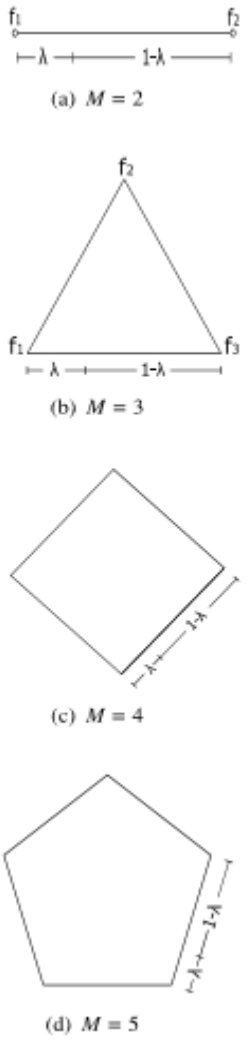

Figura 1. Enfoque poligonal para mezcla de un conjunto de funciones $\{f_1, \ldots, f_M\}$. El parámetro λ configura la mezcla de dos métodos**.**

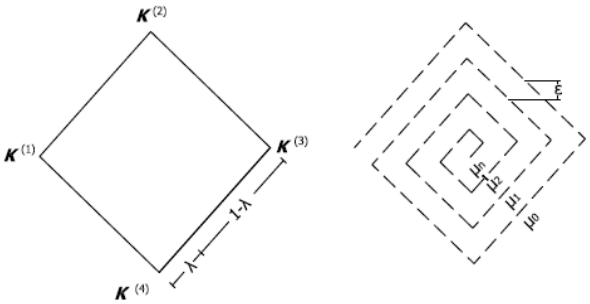

Figura 2. Homotopía geométrica para 4 métodos kernel. Este modelo permite al usuario combinar métodos seleccionando puntos dentro de la superficie del polígono. Cada punto es localizado por medio de su correspondiente par (λ, μ) el cual a su vez está asociado con un conjunto de *M* coeficientes α_m

IV. Discusión

Este trabajo busca solventar las dificultades mencionadas, siendo un puente entre el dominio de dos contextos de investigación, como lo son la RD e IV, dos campos que hacen parte del Aprendizaje de máquina (*Machine learning*), específicamente de Minería de datos (*Data Mining*) y Reconocimiento de patrones (*Pattern recognition*) y se refieren respectivamente a la representación y visualización de información cuantitativa multivariada, especialmente con un número significativamente grande de variables. Esto se puede hacer importando los conceptos de controlabilidad e interacción que están en el dominio de la IV y proyectándolos al RD para hacer un método de *pattern recognition* controlable e interactivo, ya que el objetivo de la IV, es desarrollar métodos gráficos que presenten la información más relevante para el usuario, bajo criterios de controlabilidad, donde el usuario pueda decidir cuál es el mejor modo de representar la información subyacente de sus datos en base a su objetivo de análisis, utilizando una interfaz que responda rápidamente a los cambios de parámetros, es decir, utilizar las propiedades de la visualización para hacer más legibles los resultados de la reducción de dimensión, así como más cercanos al usuario a través de combinaciones de diversos métodos de manera interactiva y amigable, de tal forma que permita la consecución gradual del objetivo en donde los pasos intermedios sean abordados en base a las teorías de la percepción humana, dando lugar a nuevos diseños de interfaces que permitan: Operaciones mentales con un rápido acceso a grandes cantidades de datos fuera de la mente, inferencia cognitiva, reducción de la demanda de la memoria de trabajo y co-participación de la máquina en una tarea conjunta, mediante el cambio gradual de las visualizaciones de forma dinámica a medida que avanza el trabajo[22].

Uno de los factores más importantes del método propuesto es la interactividad síncrona que permitirá que los métodos RD se ajusten de acuerdo al criterio del usuario, quien -aún sin conocer específicamente los métodos que se han aplicado- podrá obtener resultados confiables, involucrando un costo computacional bajo. Este trabajo podría representar un aporte en el área de Aprendizaje de máquina (*Machine learning*), y Reconocimiento de patrones (*Pattern recognition*) en términos de realizar una visualización eficiente permitiendo a un usuario, no experto o sin previo conocimiento de los métodos, obtener resultados visuales de fácil interpretación mediante el uso de una interfaz interactiva de fácil manejo que requiera de un costo computacional adecuado y que responda eficientemente a las necesidades planteadas.

V. Conclusiones

En este trabajo se presenta una metodología para el análisis visual de datos de alta dimensión en un contexto de Big Data. El objetivo de esta metodología es facilitar al usuario la tarea de seleccionar y/o sintonizar los parámetros de una técnica de visualización de una forma interactiva. En particular, en este trabajo la visualización se basa en reducción de dimensión y la interactividad está dada por la posibilidad del usuario de seleccionar los pesos o factores de ponderación de una combinación lineal de matrices kernel que representan métodos de reducción de dimensión. Como trabajo futuro, se proponer realizar interfaces interactivas y visuales que permitan evaluar el modelo.

Agradecimientos

Los autores agradecen a la Universidad Técnica del Norte, a la Corporación Universitaria Autónoma de Nariño, y a la Universidad Surcolombiana.

Referencias

[1] "¿Qué es Big Data?", 18-jun-2012. [En línea]. Disponible en: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html. [Consultado: 09-nov-2016].

[2] J. J. Camargo-Vega, J. F. Camargo-Ortega, y L. Joyanes-Aguilar, "Knowing the Big Data", *Fac. Ing.*, vol. 24, núm. 38, pp. 63–77, ene. 2015.

[3] "El Impacto de las Redes Sociales en la Propiedad Intelectual". [En línea]. Disponible en: http://www.redalyc.org/articulo.oa?id=189020164008. [Consultado: 09-nov-2016].

[4] J. C. Alvarado-PéRez, D. H. Peluffo-OrdóñEz, y R. TheróN, "Bridging the gap between human knowledge and machine learning", *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, vol. 4, núm. 1, p. 54, oct. 2015.

[5] E. Bertini y D. Lalanne, "Surveying the Complementary Role of Automatic Data Analysis and Visualization in Knowledge Discovery", en *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, New York, NY, USA, 2009, pp. 12–20.

[6] D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. 2014.

[7] J. C. Alvarado-Pérez, H. Bolaños-Ramírez, D. H. Peluffo-Ordóñez, y S. Murillo, "Knowledge discovery in databases from a perspective of intelligent information visualization", en *2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*, 2015, pp. 1–7.

[8] P. Shirley, M. Ashikhmin, S. Marschner, y T. Munzner, *Fundamentals of Computer Graphics*. CRC Press, 2009.

[9] M. F. Usama, "Mining Databases: Towards Algorithms for Knowledge Discover", vol. 21, pp. 39–48, 1998.

[10] S. Vallejos, "Minería de Datos". 2006.

[11] J. C. Riquelme, R. Ruiz, y K. Gilbert, "Minería de Datos: Conceptos y Tendencias", *Intel. Artif. Rev. Iberoam. Intel. Artif.*, 2006.

[12] P. C. Wong, "Visual data mining", *IEEE Comput. Graph. Appl.*, vol. 19, núm. 5, pp. 20–21, sep. 1999.

[13] A. Kerren, A. Ebert, y J. Meyer, Eds., *Human-centered Visualization Environments*. Berlin, Heidelberg: Springer-Verlag, 2007.

[14] Y. Wang y Q. Li, "Review on the Studies and Advances of Machine Learning Approaches", *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, núm. 2, pp. 1487–1494, feb. 2014.

[15] I. Borg y P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media, 2005.

[16] M. Belkin y P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", *Neural Comput.*, vol. 15, núm. 6, pp. 1373–1396, jun. 2003.

[17] S. T. Roweis y L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *Science*, vol. 290, núm. 5500, pp. 2323–2326, dic. 2000.

[18] G. Hinton y S. Roweis, "Stochastic Neighbor Embedding". .

[19] "Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation (PDF Download Available)". [En línea]. Disponible en: https://www.researchgate.net/publication/257352201_Type_1_and_2_mixtures_of_Kullback-

Leibler_divergences_as_cost_functions_in_dimensionality_reduction_based_on_similarity_preservation. [Consultado: 10-nov-2016].

[20] J. A. Lee, D. H. Peluffo-Ordóñez, y M. Verleysen, "Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure", *ResearchGate*, vol. 169, abr. 2015.

[21] J. Haarmann, M. P. Murphy, C. S. Peters, y P. C. Staecker, "Homotopy equivalence of finite digital images", *ArXiv E-Prints*, vol. 1408, p. arXiv:1408.2584, Agosto 2014.

[22] E. R. Tufte, *The Visual Display of Quantitative Information PAPERBACK: Second Edition PAPERBACK*. Graphics Press, 2001.