Análisis de Técnicas Inteligentes de Descubrimiento de Características para el Algoritmo Neuronal AR2P.

Ing. Ricardo Vargas
Postgrado en Computación
Universidad de Los Andes,
Mérida, Venezuela
ricardo.servitechs@gmail.com

Dr. Jose Aguilar
Tutor
CEMISID Universidad de Los Andes,
Mérida, Venezuela
aguilar@ula.ve

Msc. Eduard Puerto
Co-tutor
Grupo de Investigación GIDIS,
Universidad Francisco de Paula
Santander, Cúcuta, Colombia
eduardpuerto@ufps.edu.co

***Resumen*—Los descriptores/características que definen un problema tienen una gran importancia, porque repercuten directamente en el comportamiento de las técnicas de máquinas de aprendizaje. Los descriptores influyen en la calidad de los algoritmos de aprendizaje, ya sea en procesos de reconocimiento de patrones, descubrimiento de patrones, etc. En ese sentido, recientemente se ha creado un área denominada ingeniería de descriptores, que estudia el proceso completo que se debe seguir para obtener estos descriptores, incluyendo las posibles técnicas que utilizarán en los distintos contextos que se pueden presentar. Es por esto que en esta tesis se propone realizar un estudio exhaustivo sobre el proceso de ingeniería de descriptores para el algoritmo neuronal AR2P, analizando las distintas técnicas que se podrían utilizar, específicamente para el tratamiento de imágenes, video, stream de datos y patrones dinámicos, de tal manera de proponer un esquema genérico de tratamiento de características para el algoritmo neuronal AR2P.**

Palabras Clave—Ciencia de los Datos, Ingeniería de Descriptores/Características, Algoritmo Neuronal Ar2p, Aprendizaje Neuronal.

I. Introducción

En el área de aprendizaje de máquina, las diferentes técnicas (supervisadas, semi-supervisadas o no supervisadas) son altamente sensibles a los datos de entrenamiento. Por lo tanto, uno de los elementos fundamentales a tener en cuenta son las características (también denominado descriptores o propiedades) que se usan para describir el fenómeno en estudio. Descubrir estos descriptores del problema estudiado es de suma importancia, ya que la calidad del modelo de aprendizaje a construir depende de ellos. Además, estos descriptores afectan también la eficiencia (o complejidad computacional) de los algoritmos de aprendizaje. En la literatura, al proceso de descubrir descriptores se le ha ido denominando ingeniería de descriptores [1, 2, 3]. En general, la ingeniería de descriptores consisten en tres grandes fases para el descubrimiento de los descriptores: *la fase de extracción de características*, en la cual se descubren y definen las características de base del conjunto de datos recopilados, que describen adecuadamente al fenómeno estudiado; la *fase de construcción de características*, en la cual se usan al conjunto de datos y sus características de base para definir nuevos descriptores que reflejen propiedades específicas en los datos recopilados, aplicando técnicas de inferencia, cálculo estadístico, entre otras, de tal manera de enriquecer la información sobre el sistema bajo estudio; y por último, *la fase de selección de características* donde se aplican técnicas que buscan seleccionar o reducir el conjunto de *características/descriptores* a utilizar, descartando los redundantes y/o innecesarios, quedándose así solo con el conjunto de características ideales para el problema dado.

Por otro lado, AR2P [12, 13] es un algoritmo de reconocimiento de patrones basado en teorías sobre el funcionamiento del neocórtex del cerebro humano, como la teoría de la mente para el reconocimiento de patrones (PRTM), la cual describe el proceso de reconocimiento de patrones que se lleva a cabo en el neocórtex de la siguiente manera: el cerebro se organiza en módulos conformados por unas 100 neuronas, y estos módulos están organizados en niveles, reflejando así una organización jerárquica del mismo. Partiendo de ese modelo, es posible desarrollar procesos de reconocimiento de patrones de manera recursiva, tal que patrones de gran complejidad son compuesto por patrones más simples, y así sucesivamente. De esta manera, al descubrirse los patrones de los niveles atómicos de un patrón dado, permite su reconocimiento. Viceversa, para reconocer un patrón complejo debe primero ser descompuesto en sus patrones más simples, y así sucesivamente. Finalmente, para el algoritmo AR2P se han desarrollado algoritmos de aprendizaje supervisado y profundo en trabajos previos [13, 14].

II. Formulación del problema

La calidad de los modelos de conocimiento desarrollados por las Máquinas de Aprendizaje, como los modelos de reconocimiento, los modelos de clasificación o los modelos de agrupamiento, dependen de varios factores, pero uno de los más importantes es el conjunto de descriptores de base utilizados. En particular, el algoritmo AR2P, por su enfoque jerárquico, tiene una alta dependencia de los patrones atómicos utilizados en sus procesos de reconocimiento, los cuales son los descriptores de base del patrón a reconocer.

Para el algoritmo AR2P se han desarrollado algoritmos de aprendizaje supervisado y profundo en trabajos previos [13, 14], pero se debe mejorar los procesos de identificación automática de los descriptores atómicos de los patrones a reconocer. Es allí donde este trabajo pretende hacer su aporte. En ese sentido, esta tesis busca responder al siguiente interrogante ¿Qué técnica(s) de descubrimiento de características desde el campo de Máquinas de Aprendizaje permite mejorar la definición de descriptores para AR2P?

En particular, esa pregunta es muy pertinente en el algoritmo de aprendizaje profundo para AR2P, el cual está basado en un esquema de aprendizaje semi-supervisado. En ese sentido, el algoritmo de aprendizaje profundo está compuesto de varias fases, pero la que tiene que ver con la pregunta anterior es la fase de descubrimiento de patrones atómicos, la cual debe ser resuelta por un algoritmo como el que se pretende proponer en este trabajo. Para ello, durante este trabajo se analizarán varias técnicas de descubrimiento de descriptores basadas en Máquinas de Aprendizaje, para adecuar una para el proceso de reconocimiento de patrones realizado con AR2P.

III. Antecedentes

El proceso de ingeniería de descriptores/características ha sido descrito en una serie de trabajos [1, 2, 3] en las cuales se destacan los pasos que la componen (extracción, construcción y selección de características). En dichos pasos, diferentes

técnicas se pueden aplicar. Además, otros trabajos aplican procesos simples de análisis de descriptores usando técnicas para determinar los descriptores que mejor representan el problema bajo estudio, como k-means, la cual es una técnica de agrupamiento (clustering) que debido a su rapidez y fácil implementación ha sido usada para extraer representaciones de imágenes a gran escala [5], o el análisis de componentes principales (PCA) para encontrar características importantes del problema a analizar, como en [4] para el problema de reconocer la identidad urbana en espacios públicos.

Algunos de los trabajos de extracción de características recientes son [15, 16]. En [15] se propone un enfoque para reconocimiento facial bajo condiciones de variación de luz. Este enfoque consta de tres fases, la primera en la cual se realiza una normalización de los componentes de iluminación con métodos como el "Weber-face", en segundo lugar, se realiza la extracción de descriptores Para esto, los autores establecen la diferencia entre dos posibles grupos de técnicas: técnicas holísticas y técnicas locales. Entre las técnicas holísticas se encuentran PCA o LDA, y entre las locales métodos como patrones binarios locales (LBP) o patrones direccionales locales (LDP) la cual es la usada en ese enfoque propuesto. Finalmente, en la tercera fase usan máquinas de soporte vectoriales (SVM) para realizar la clasificación. Los resultados son comparados con otros enfoques compuestos por diferentes métodos en cada fase. Como complemento a el área de extracción de características en imágenes, en [10] se realiza una descripción detallada sobre este proceso profundizando en las diferentes técnicas existentes, tipos de descriptores que pueden existir, etc. En [16] se propone un proceso de extracción de características para minería de texto, en el cual divide el proceso en una extracción a nivel sintáctico y otra a nivel semántico. A nivel sintáctico proponen una versión de $\chi 2$ (Chi-squared statistics) mejorada (ICHI) aplicada sobre una matriz de palabras para tratar con los datos desbalanceados, y a nivel semántico una técnica llamada asignación latente de Dirichlet sobre una representación de documentos por tópicos en lugar de espacios de palabras. Finalmente, se realiza una fusión serial sobre ambos resultados para obtener el conjunto de características final.

En cuanto a propuesta de construcción de características, el trabajo [17] propone un método basado en programación genética (GP) para obtener características faltantes en datos incompletos. En esta técnica se utilizan funciones de intervalos como parte del conjunto de funciones del GP, resultando en una sustitución del valor faltante por un intervalo de posibles valores, o si es un valor no faltante se sustituye por un intervalo donde tanto el límite superior como inferior tienen el valor original. En [18] se propone un proceso de construcción de características basadas en entropía, para la detección de ataques de DDoS. Este proceso consta de tres partes: primero, se extraen características crudas de cada cabecera de los paquetes de red. Datos como direcciones IP origen y destino, puertos origen y destino y protocolos, son los más comúnmente usados. En segundo lugar, se calcula la entropía de Shannon usando un intervalo de tiempo específico, y por último, se construyen nuevas características con variaciones de entropía, usando combinaciones de características generadas en los pasos previos. Finalmente, en selección de características algunos trabajos son: En [19] se propone un enfoque de selección de características para previsión de precios y cargas en sistemas de suministro eléctrico. Este enfoque mezcla características de modelos de filtro (Filtering) y envoltorios (Wrapping). De este modo, el método propuesto selecciona un subconjunto de características tomando en cuenta criterios como relevancia, redundancia e interacción entre características, de forma independiente del algoritmo de aprendizaje que se usará. En [20] se proponen mejorar la identificación de patologías del pecho mediante el análisis de radiografías. En este enfoque se extraen características de estas imágenes radiológicas usando redes neuronales convolucionales, luego establecen una fase de reducción de características realizando una combinación de ellas por medio de una técnica de fusión lineal ponderada basada en los pesos de probabilidad de cada clase, y una fase de selección de características en la cual usando pruebas de Kruskal-Wallis sobre la varianza de los datos, determinar las características más informativas. Por último, una técnica de gran importancia en este proceso es la de Random Forest, en [11] se emplea esta técnica para establecer la importancia de las variables que pueden ser usadas en procesos de detección de fallos. Esta técnica consiste en generar un conjunto de árboles de clasificación y regresión, cada uno usando aleatoriamente un subconjunto de las características disponibles como nodos, y se van asignando pesos a cada característica de acuerdo a qué tan relevante fue en el resultado del árbol de decisión al cual pertenecía.

Otro ámbito importante actual es el referente al análisis de datos en línea o de serie temporal (stream de datos), en los cuales no se dispone de los datos de forma inmediata, y por lo tanto el proceso para determinar los descriptores se debe realizar mediante cálculos incrementales, aplicando técnicas como la media móvil (moving average), o el aprendizaje incremental, para trabajar con estos datos dependientes del tiempo [6, 7, 8, 9]. Por ejemplo, en [8] se realiza un estudio exhaustivo sobre la media móvil, y propone un marco de trabajo para su definición y desarrollo, tomando en cuenta diferentes métodos con diferentes parámetros y definiciones de métricas, por ejemplo, adaptando métodos para tomar en cuenta el espacio temporal entre los datos recibidos. Además, en este marco de trabajo se incluyen también los conceptos de histogramas de movimiento y medidas de frecuencia que facilitan las aproximaciones de cantidades dependientes del tiempo. En [7] se propone un modelo generalizado de media móvil exponencial (EMA por sus siglas en inglés) para realizar predicciones en mercados financieros. El modelo además combina estas técnicas con conceptos del área de aprendizaje de máquina para tomar ventaja de la no linealidad que presentan algunos de estos métodos, necesarios para realizar las predicciones, mediante un algoritmo de estimación secuencial con detección de anomalías, un modelo de promediado (model averaging), y modelos de retornos de activos basados en series de tiempo.

Finalmente, el algoritmo AR2P es un algoritmo de reconocimiento de patrones que utiliza la idea de jerarquía neocortical y de desagregación/integración del patrón en el proceso de reconocimiento [12]. Este algoritmo se basa en las teorías que describen el funcionamiento del neocórtex humano que explotan la idea de recursividad. En dicho algoritmo se han desarrollados Algoritmos de Aprendizaje Supervisado [13] y de Aprendizaje Profundo [14]. El algoritmo de Aprendizaje Supervisado está compuesto por dos mecanismos, uno llamado Aprendizaje_nuevo, en el cual se aprenden nuevos patrones creándose sus nuevos módulos de reconocimiento, y el segundo llamado Aprendizaje_por_refuerzo, para reforzar los patrones ya aprendidos y adaptar sus módulos a los cambios presentados en el mismo. Este algoritmo ha sido probado en diversos contextos como reconocimiento de textos e imágenes [12, 13]. En cuanto al algoritmo de aprendizaje profundo [14], se extienden sus capacidades de aprendizaje supervisado con un esquema de aprendizaje semi-supervisado. Este enfoque está compuesto por tres fases, la primera fase de descubrimiento de los descriptores atómicos, la segunda de agregación, creando una jerarquía con estos descriptores, y la tercera de clasificación usando los descriptores generados en la segunda

fase. Este es un esquema semi-supervisado, ya que las dos primeras fases usan un aprendizaje supervisado, mientras que la tercera es no supervisado.

IV. Objetivos

A. General

Implementar técnicas de descubrimiento de características/descriptores óptimos para el proceso de reconocimiento de patrones del algoritmo neuronal AR2P.

B. Específicos

1) Realizar un estado de arte sobre las técnicas de descubrimiento de características/descriptores.

2) Identificar las posibles técnicas usadas para el descubrimiento de características/descriptores para AR2P.

3) Generar un prototipo funcional adaptado a AR2P, que genere los descriptores óptimos para el proceso de reconocimiento de patrones.

V. Metodología

El proceso de desarrollo de la investigación consta de 4 fases.

Fase 1: Descripción de los procesos necesarios para el descubrimiento de características. Esta fase analizará los pasos de extracción, construcción y selección de características que componen el proceso de descubrimiento.

Fase 2: Analizar el proceso de selección de técnicas de descubrimiento de características, y su relación con el algoritmo neuronal AR2P.

Fase 3: Desarrollar el algoritmo de descubrimiento de características adecuado para el algoritmo neuronal AR2P.

Fase 4: Realizar pruebas del algoritmo desarrollado sobre distintos conjuntos de datos (imágenes, stream de datos, objetivos dinámicos, entre otros) y escenarios.

VI. Resultados Esperados

Un modelo de descubrimiento de características para el algoritmo neuronal AR2P, con su Prototipo funcional, además de escenarios experimentales de prueba.

VII. Programación de Actividades

El plan de actividades se muestra en el diagrama de Gantt de la figura 1.

Referencias

[1] H. Motoda and H. Liu “Feature Selection, Extraction and Construction”, Communication of Institute of Information and Computing Machinery, Vol. 5, pp. 67–72, 2002.

[2] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh, “Feature Extraction Foundations and Applications”, Studies in Fuzziness and Soft Computing, Vol. 207, 2006.

[3] S. Wang, J. Tang, H. Liu, “Encyclopedia of Machine Learning and Data Mining”, Second Edition, Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, Chapter Feature Selection pp. 878-882, 2008.

[4] M. Chang, P. Buš, G. Schmitt “Feature Extraction and K-means Clustering Approach to Explore Important Features of Urban Identity”, 16th IEEE International Conference on Machine Learning and Applications, 2017.

[5] A. Coates, A. Y. Ng “Learning Feature Representations with K-means” G. Montavon, G. B. Orr, K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade, 2nd edn, pp 561-580, Springer LNCS 7700, 2012.

[6] G. R. Arce, “Nonlinear Signal Processing”, John Wiley & Sons, Inc., pp. 81-250, 2005.

[7] M. Nakano, A. Takahashi, and S. Takahashi, “Generalized exponential moving average (ema) model with particle filtering and anomaly detection”, Expert Systems with Applications, Vol. 73, pp. 187 – 200, 2017.

[8] M. Menth and F. Hauser, “On moving averages, histograms and time-dependent rates for online measurement”, 8th ACM/SPEC on International Conference on Performance Engineering, pp. 103–114, 2017.

[9] G. E. P. Box and G. Jenkins, “Time Series Analysis, Forecasting and Control”, Holden-Day, Incorporated, 4th edn, pp. 551-595, 1990.

[10] G. Kumar, P. K. Bhatia “A Detailed Review of Feature Extraction in Image Processing Systems”, Fourth International Conference on Advanced Computing & Communication Technologies, 2014.

[11] C. Aldrich, L. Auret “Fault detection and diagnosis with random forest feature extraction and variable importance methods”, 13th Symposium on Automation in Mining, Mineral and Metal Processing Cape Town, 2010.

[12] P. Eduard, J. Aguilar “A Recursive Pattern Recognition Algorithm”, Rev. Téc. Ing. Univ. Zulia. Vol. 40, No 2, pp. 95-104, 2017.

[13] P. Eduard, J. Aguilar “Learning Algorithm for the Recursive Pattern Recognition Model”, Applied Artificial Intelligence, Vol. 30, No 7, pp. 662-678, 2016.

[14] P. Eduard, J. Aguilar, J. Reyes, D. Sarkar “A Deep Learning Architecture for the Recursive Patterns Recognition Model”, Enviado a Publicación, 2018.

[15] Chi-Kien Tran, Chin-Dar Tseng, Pei-Ju Chao, Chin-Shiuh Shieh, Liyun Chang, Tsair-Fwu Lee ”Face Recognition under Varying Lighting Conditions: A Combination of Weber-face and Local Directional Pattern for Feature Extraction and Support Vector Machines for Classification”, Journal of Information Hiding and Multimedia Signal Processing, Vol. 8 No 5, pp. 1009-1019, 2017.

[16] F. Wang, T. Xu, T. Tang, M. Zhou, H. Wang “Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems”, IEEE Transactions on Intelligent Transportation Systems, Vol. 18, No. 1, pp. 49-58, 2017.

[17] C. Tran, M. Zhang, P. Andreae, B. Xue “Genetic Programming based Feature Construction for Classification with Incomplete Data”, GECCO '17 Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1033-1040, 2017.

[18] A. Koay, A. Chen, I. Welch, W. K. G. Seah “A New Multi Classifier System using Entropy-based Features in DDoS Attack Detection”, International Conference on Information Networking (ICOIN), 2018.

[19] O. Abedinia, N. Amjady, H. Zareipour “A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems”, IEEE Transactions on Power Systems, Vol. 32, No. 1, 2017.

[20] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, H. Greenspan “Chest pathology identification using deep feature selection with non-medical training”, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 6, No. 3, pp. 259-263, 2016.

Nombre	Trab...
Finalizar propuesta	5d
Realizar estado del arte	3d
Proceso de descubrimiento	**10d**
Establecer el proceso de descubrimiento de características	3d
Determinar técnicas	2d
Desarrollar modelo de descubrimiento	5d
Pruebas	**6d**
Realizar pruebas sobre datasets	3d
Realizar pruebas sobre ar2p	3d
Generar manuscrito	5d

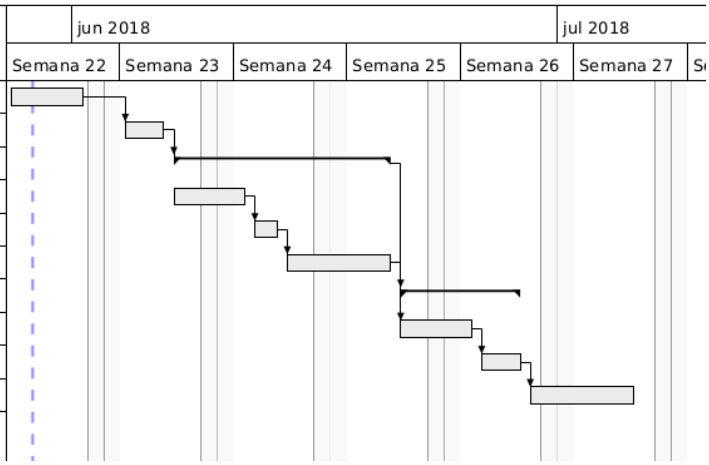

Fig. 1. Planificación de actividades