AUTOMATIZACIÓN DEL APRENDIZAJE DE MÁQUINAS ("AUTOML"): LOGROS Y OBSTÁCULOS

MACHINE LEARNING AUTOMATION ("AUTOML"): CURRENT ACHIEVEMENTS AND OBSTACLES

Investigadores USAL: Bender, Adrián (bender.adrian@usal.edu.ar); Nicolet, Santiago; Macrino, Matías.

Palabras clave: AutoML; Aprendizaje Automático; Automatización; Simulación; Minería de Datos.

Keywords*: AutoML; Machine Learning; Automation; Simulation; Data Mining*

Resumen

La Automatización del Aprendizaje de Máquinas (AutoML) pretende complementar o simular la tarea de los expertos en Aprendizaje Automático en el desarrollo de un proceso de Minería de Datos. El avance reciente en el Aprendizaje Automático y las ventajas competitivas que permite el descubrimiento de conocimiento en los datos generan un auge por el desarrollo de aplicaciones que automaticen las tareas de este flujo de trabajo.

Es por eso que existe una creciente comunidad generada en torno a la creación de herramientas que automatizan estas tareas, cuyo éxito hoy día depende fundamentalmente de expertos en *Machine Learning*, quienes preprocesan los datos, construyen los modelos eligiendo los algoritmos apropiados y configuran sus hiperparámetros.

Diversas son las técnicas con las cuales se pretende automatizar estas actividades, que en el caso de los expertos humanos es llevada a cabo con conocimiento, intuición, juicio y razonamiento.

Un tema a considerar en la evaluación de estas técnicas es que la productividad en la Minería de Datos no es una cuestión cuantitativa, es más bien un problema de la calidad de lo que los procesos produzcan. En el contexto de conocimiento de los datos, la calidad se refiere a la validez y relevancia de los patrones que los modelos pueden descubrir a partir de los datos. Entonces, será interesante saber:

¿Qué sucederá cuando se automatice el trabajo de todos estos expertos?

¿Qué pasará con la calidad cuando se "democratice" aún más el campo del Aprendizaje Automático proporcionando a cualquier persona las herramientas de análisis automático?

Este proyecto de investigación se propuso como objetivo general el relevamiento del alcance y de la eficacia de las herramientas de AutoML disponibles, buscando que los resultados permitieran contribuir al conocimiento del estado del arte de esta incipiente área, cuantificar el grado de eficacia que tienen las herramientas existentes e identificar áreas de mejora para la automatización de esta ciencia que ha cobrado tanta importancia recientemente.

Para ello, se desarrolló un sistema de métricas que permitiera relevar el alcance y las capacidades de automatización de los *frameworks*. Se seleccionaron los de mayor uso y se identificaron conjuntos

de datos a ser utilizados como muestra de los problemas que estas herramientas permiten resolver. Se efectuó una prueba comparativa que evaluara la performance para la resolución de los problemas seleccionados.

Como resultado del trabajo, se elaboró el documento "Evaluación Comparativa de Herramientas AutoML de Código Abierto", enviado para su evaluación al CoNaIISI 2019 – 7. ° Congreso Nacional de Ingeniería Informática – Sistemas de Información. El *paper* fue aprobado y presentado en dicho congreso internacional. Allí se contrastaron los valores obtenidos por las tres herramientas evaluadas y por la línea base establecida. El análisis sobre estos permitió concluir que todas las herramientas mostraron cierto nivel de eficacia y que lograron mejores resultados que los que un usuario obtendría de forma básica. El trabajo incluyó una segunda instancia de evaluación, y allí las diferencias no resultaron significativas. Afirmamos que entre sus posibles causas podría estar la sobreadaptación de los pipelines generados, lo cual consideramos una interesante línea de investigación futura. Al contrastar los resultados de las herramientas entre sí, no observamos diferencias significativas entre ellas. El trabajo también incluyó un detalle de la experiencia con dichas herramientas, y entre las conclusiones se mencionaron las principales características de cada una: la generación de modelos ensamblados y el módulo de inicio rápido de Auto-sklearn, la posibilidad de exportación del modelo de TPOT para ser utilizado sin dependencias, y la facilidad de uso de Auto-WEKA, que permite obtener buenos resultados con un solo clic.

Abstract

Machine Learning Automation (AutoML) aims at simulating or complementing the work of Machine Learning experts in developing Data Mining processes. Recent developments in Machine Learning and the competitive advantages in data knowledge generate a hype in the development of applications that automate the tasks of this workflow.

That is why there is a growing community around the creation of tools that automate these tasks. Nowadays, success depends fundamentally on experts in Machine Learning, who pre-process the data, build the models by choosing the appropriate algorithms, and configure their hyperparameters.

Various techniques are used to automate these activities, which, in the case of human experts, is carried out with knowledge, intuition, judgment and reasoning.

An issue to consider while evaluating these techniques is that productivity in Data Mining is not a quantitative issue, but rather a problem of the quality of what the processes produce. In the context of data knowledge, quality refers to the validity and relevance of the patterns that models can discover from the data. It will be interesting to know:

** What will happen when the work of all these experts is automated?*

** What will happen to quality when the "democratization" provides machine analysis tools to anyone?*

The general purpose of this research project is to survey the scope and effectiveness of the available AutoML tools. It seeks to contribute to the knowledge of the state of art of this emerging area. It also quantifies the degree of effectiveness of the existing tools and identifies areas for improvement in the automation of this incipient but important science.

To achieve these objectives, a system was developed which let us measure the scope and automation capabilities of the studied frameworks. The most widely used tools were selected. We selected special data sets to be used as samples of the problems that these tools allow us to solve. A comparative test was performed evaluating the performance to solve the selected problems.

As a result of the work, the document "Comparative Evaluation of Open Source AutoML Tools" was prepared and sent for evaluation to CoNaIISI 2019 - 7th National Congress of Computer

Engineering - Information Systems. The paper was approved and was presented at that international congress. There, the values obtained by the three evaluated tools and by the established baseline were compared. Their analysis allowed us to conclude that all the tools displayed a certain level of effectiveness, achieving better results than what a user would do in a basic way.

The work included a second data instance of evaluation, and the differences were not significant. A possible cause could be the over-adaptation of the pipelines, which we consider an interesting line for future research. When contrasting the results of the tools with each other, we did not observe significant differences between them. The work also included a detailed usage experience with these tools. In addition to the conclusions, we mentioned the main characteristics of each tool: the generation of assembled models and the Auto-sklearn quick start module, the possibility of exporting the TPOT model to use it without dependencies, and the easy use of Auto-WEKA, which allows to obtain good results with a single click.