Machine Learning para Ciencias Sociales (Introducción)

Pavel Coronado [1]

[1] Departamento de Economía y QLAB - PUCP

January 13, 2025

Indice

Inteligencia Artificial

- En los inicios de la Inteligencia Artificial (IA), ésta resolvia problemas que eran dificiles para humanos pero relativamente faciles para computadores, problemas que podian ser descritos por unalista de reglas formales y matematicas.
- El reto para la inteligencia artificial probó ser el de resolver problemas que eran sencillos para humanos, problemas que para nosotros se resuelven casi intuitivamente como reconocer voces o rostros en una imagen.
- Enfoques alternativos de la IA buscaban colocar conocimiento acerca del mundo en lenguajes formales de tal manera que las computadoras pudieran "razonar" usando reglas de inferencia lógica. Este enfoque es conocido como **el enfoque basado en conocimiento**. Uno de los proyectos mas famosos de este enfoque es el Cyc (Lenat and Guha, 1989). Cyc falló al entender una historia acerca de una persona llamada Fred que se estaba afeitando. Puede ver una entrevista a James Lenat sobre avances recientes en este enfoque aquí.

ELSEVIER
Artificial Intelligence
Volume 61, Issue 1, May 1993, Pages 41-52

Book review

Building large knowledge-based systems: Representation and inference in the cyc project: D.B. Lenat and R.V. Guha ☆

Charles Elkan **, Russell Greiner †

Show more

+ Add to Mendeley   Share   Cite

https://doi.org/10.1016/0004-3702(93)90092-P   Get rights and content

Abstract

The book under review here, *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*, describes progress so far in an attempt to build a system that is intended to exhibit general common-sense reasoning ability. This review first discusses aspects of the Cyc system, with a

Inteligencia Artificial

- Lenat and Guha believe that the key to dealing with unanticipated situations as successfully as humans is to possess factual information and reasoning facilities as comprehensive as those possessed by humans.
- The Cyc project therefore addresses the tremendous task of codifying a vast quantity of knowledge about the world, what the authors call "consensus reality"
- The Cyc team is attempting to capture the background knowledge possessed by a typical late twentieth century inhabitant of the United States.
- The goals of the Cyc project go beyond just making this mass of information available for retrieval. Cyc therefore will include a wide range of reasoning facilities, including procedures for general deduction and analogical inference.

Machine Learning

- Machine Learning (ML) es una rama de la Inteligencia Artificial (IA).
- El enfoque de la Inteligencia Artificial que ha tenido rotundos exitos en los ultimos tiempos es un enfoque que trata que las "maquinas" aprendan su propio conocimiento de los datos, este es el **el enfoque basado en aprendizaje o machine learning**
- Es considerado como parte de la inteligencia artificial porque implica la creación de "dispositivos" que **aprendan** a encontrar patrones en los datos con escaza o nula ayuda humana y que luego sirvan para resolver tareas normalmente hechas por humanos.
- Definición Formal: "Un algoritmo o programa de computadora aprende de la experiencia E con respecto a alguna tarea T, si su performance al realizar la tarea T, mejora con la experiencia E" (Mitchel, 1997)
- Definición Informal: Algortimos que mejoran al realizar una tarea con la experiencia.
- Como se verá, tienen que ver mas con estadistica y optimización que con lógica.
- Machine Learning cuenta con aplicaciones en negocios, medicina, bioquímica y, por supuesto, en las ciencias sociales.

Machine Learning. Un poco de Historia

- 1950: Alan Turing propuso el famoso Test de Turing
- 1952. Artur Samuel."Campo de Estudio que da a las computadoras la capacidad de aprender sin ser explicitamente programadas"
- 1952. Arthur Samuel creó un programa capaz de derrotar a su propio creador y al campeón mundial de "damas". El programa era un Algoritmo Shanon mini-max.

A. L. Samuel

Some Studies in Machine Learning Using the Game of Checkers

Abstract: Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these experiments are, of course, applicable to many other situations.

PUCP

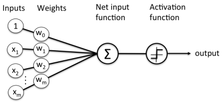
Inteligencia Artificial

Machine Learning-Definición

Aplicaciones en CCSS

Machine Learning. Un poco de Historia

- 1958:Frank Ronsenblatt diseñó la primera aritificial neural network (The perceptron)



- 1986: Geoffrey Hinton introduce el concepto de "backpropagation" el cual permitió grandes avanzacesen Redes Neuronales artificiales.

- 1994: Gerry Tesauro enseño a un red neuronal a jugar "backgammon". La maquina jugo 200 mil veces contra sí misma y luego derrotó al campeón mundial.

Machine Learning. Un poco de Historia

- 1980s: Yan le Cunn presenta las Convolutional Neural Nets
- 1997: Deep Blue derrota a Garry Kasparov
- 1999: Diagnosticos de computadora detectan mas canceres que los doctores.
- 2012: AlexNet ganó la competencia de ImageNet, lo cual llevó al uso de GPUs y al uso de Convolutional Neural Nets en Machine Learning (Alex Krizhevsky, Ilya Sutskever)

Machine Learning. Un poco de Historia

- 2014: Se crea el mecanismo de atención. Bengio et al (2014) "Neural Machine Translation by Jointly Learning to Align and Translate"

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ABSTRACT

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

Machine Learning. Un poco de Historia

- 2017: Cientificos de Google crean la red transformer. Vaswani et al. (2017) "Attention is All you Need"

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Machine Learning. Un poco de Historia

- 2019: LeCun, Hinton y Bengio ganan el Premio Turing

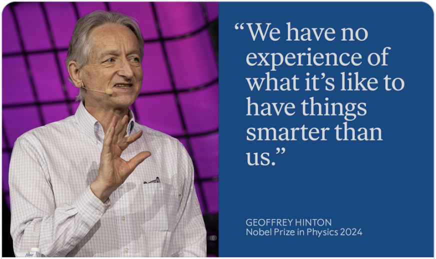
- 2022: Lanzamiento del Chat GPT. Sobre el futuro de la IA puede ver las opiniones de Geoffrey Hinton aquí.

Machine Learning. Un poco de Historia

- Y. Le Cun no cree que los riesgos sean importantes.

Yann LeCun
January 8 at 3:54 PM ·

That's just false.
Most of us are intimately familiar with interacting with people who are smarter than us.
Most scientists actively seek to work with people who are smarter than them in at least a few domains. It amplifies their intellectual power.
It will be the same with AI assistants: they will amplify our intellectual power.



Machine Learning. Un poco de Historia

- Bengio: Sí existen riesgos potenciales

Yoshua Bengio
I stand with Geoff Hinton. He is right: we have no experience of dealing with non-human entities that are significantly smarter than us. **Yann** often says that we will not build them if we are not sure that they will act morally. The problem is that (a) we don't know how to do that and yet we race ahead (b) it may well be that some people won't be sufficiently cautious, especially if there are no rules and norms to minimize those risks (c) there are people who seem to welcome the idea of superintelligent AIs taking over the control of our planet and (d) there are arguments that I find theoretically very convincing for the unintentional emergence of a self-preservation goal in RL agents, which does not mean such entities will be immoral, only that they will value their survival more than ours (and we would also probably choose to terminate an entity which could kill us to protect itself, rather than take a chance with our lives).

2d Like Reply 44

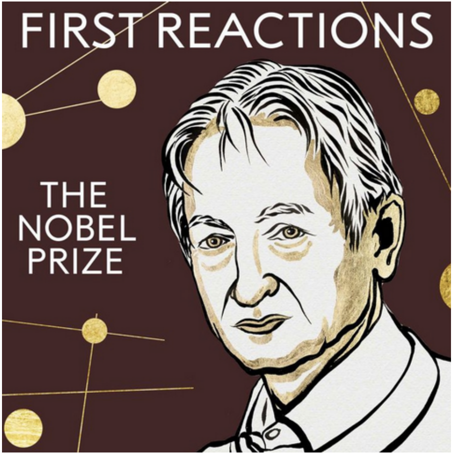
Machine Learning. Un poco de Historia

- Bengio: Sí existen riesgos potenciales

> **Yoshua Bengio**
> I don't think we can have a precise probabilistic estimate of our chances of survival. From my informal discussions with Geoff, he is more pessimistic than I am. The reason I am maybe more optimistic is that I think that we could build safe superintelligent AI, but we have to do it in full conscience (and not denial) of the risks, so that we can avoid the traps that may otherwise await us. I am writing a paper where I summarize these ideas.
>
> 2d Like Reply 12

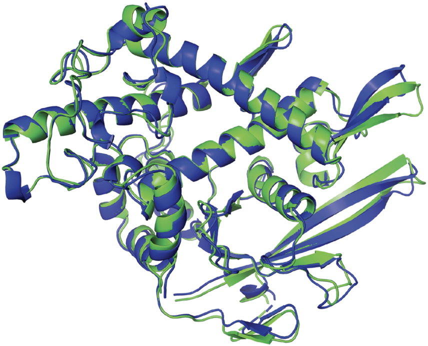
Machine Learning. Un poco de Historia

- Hinton Won Nobel Prize in Physics. First reaction aquí.



Machine Learning. Un poco de Historia

- Nobel Prize in Chemistry. Baker,Jumper and Hassabi. Mas información aquí aquí.



Structures of a protein that were predicted by artificial intelligence (blue) and experimentally determined (green) match almost perfectly.

La clave es aprendizaje de Patrones y Estructuras

- Aprendemos a traves de ejemplos o de casos positivos y negativos
- Aunque muchas veces no seamos del todo conscientes, nosotros aprendemos patrones.
- Las "maquinas" pueden aprender patrones que nosotros no podemos identificar.

Aprendemos por Experiencia

- ¿ Cómo aprendemos a Identificar un perro?
- A partir de la instrucción de nuestros padres
- a partir de ver muchos perros siendo llamados perros
- A partir de la identificacion de caracteristicas particulares

Aprendemos Patrones


Memory: The Key To Chess?

- Gran Maestro (GM) puede reconocer los juegos por fechas y oponentes
- GM reconocer muchas mas piezas en un tablero de un juego que persona promedio
- GM no es mucho mejor que persona promedio para recordar posiciones alocadas aleatoriamente.
- Según Herbert Simon esto se debe a que GMs aprenden patrones en bloque comunes en los juegos (chunks)

Conclusiones

- Mucho de nuestro propio aprendizaje ocurre a traves de la experiencia.
- Este aprendizaje consiste en recolectar mucha información.
- El aprendizaje consiste también en detectar patrones en la realidad para cada tipo de tarea.
- Pueden estos patrones ser identificados por Máquinás? Sí.
- In a nutshell: pueden representarse por funciones $f(x)$. Estas funciones pueden se de una gran complejidad.

¿Cómo Modelamos estos Patrones?

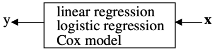
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

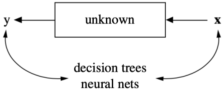
Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

¿Cómo Modelamos estos Patrones?



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

(a) Data Modeling Culture



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

(b) Algorithmic Culture

Aplicaciones en CCSS

- Breiman (2001) If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.
- “Muchos problemas relevantes en ciencias sociales son eminentemente problemas de prediccción y por tanto requieren principalmente estimar su error de predición o bondad de ajuste. Mullainathan y Spiess(2017) JEL
- “Economistas y econometristas necesitan ir mas allá de las herramientas econometricas tradicionales y adoptar un nuevo set de herramientas” Athey y Imbens (2019). Annual Review of Economics. AEA.
- Los fundamentos y los métodos más comunes de machine learning son relativamente faciles de aprender solo con una base en estadística y en análisis de regresión multivariado. Por tanto, esta al alcance de muchos estudiantes de ciencias sociales.

Machine Learning e Inferencia causal

Objetivo de ML

Minimizar la función de pérdida: $\hat{f}$ tal que resuelva

$$min_{f \epsilon \mathbb{F}} \mathbb{E}[l(f(x), y)]$$

usar los datos para escoger una función que predice bien en "nuevas observaciones". Vapnik(1999) muestra que este problema tiene solución (Empirical Risk Minimization Principle). EL ERM, usa el principio de consistencia universal y la ley de los grandes números.

Objetivo Inferencial Causal

Encontrar un parametro θ tal que resuelva:

$$min_{\theta} \mathbb{E}[l(f(x, \theta), y)]$$

y además $\mathbb{E}(\theta) = \theta$.

Funciones de Perdida or Loss Function en ML

Regression Loss ó Perdida en Regresión

- Pérdida Cuadratica: $l(f(x), y) = (y - f(x))^2$
- Pérdida Absoluta: $l(f(x), y) = |y - f(x)|$

Classification Loss o Pérdida en Clasificación

- 0-1 loss, error de clasificacion
- hinge loss $l(f(x), y) = max(0, 1 - yf(x))$

Objetivo de Minimización

En ML buscaremos optimizar:

$$\mathbb{E}[l(f(x), y)]$$

o en una muestra en particular:

$$R_{emp} = \frac{1}{n} \sum_i l(f(x), y)$$

Esta última es llamado Empirical Loss o Pérdida empírica (en la muestra)

Aplicaciones en CCSS

Scaling to Very Very Large Corpora for Natural Language Disambiguation

Michele Banko and Eric Brill
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
{mbanko,brill}@microsoft.com

Abstract

The amount of readily available on-line text has reached hundreds of billions of words and continues to grow. Yet for most core natural language tasks, algorithms continue to be optimized, tested and compared after training on corpora consisting of only one million words or less. In this paper, we evaluate the performance of different learning methods on a prototypical natural language disambiguation task, confusion set disambiguation, when trained on orders of magnitude more labeled data than has previously been used. We are fortunate that for this particular application, correctly labeled training data is free. Since this will often not be the case, we examine methods for effectively exploiting very large corpora when labeled data comes at a cost.

1 Introduction

Machine learning techniques, which automatically learn linguistic information from

potentially large cost of annotating data for those learning methods that rely on labeled text.

The empirical NLP community has put substantial effort into evaluating performance of a large number of machine learning methods over fixed, and relatively small, data sets. Yet since we now have access to significantly more data, one has to wonder what conclusions that have been drawn on small data sets may carry over when these learning methods are trained using much larger corpora.

In this paper, we present a study of the effects of data size on machine learning for natural language disambiguation. In particular, we study the problem of selection among confusable words, using orders of magnitude more training data than has ever been applied to this problem. First we show learning curves for four different machine learning algorithms. Next, we consider the efficacy of voting, sample selection and partially unsupervised learning with large training corpora, in hopes of being able to obtain the benefits that come from significantly larger training corpora without incurring too large a cost.

2 Confusion Set Disambiguation

Confusion set disambiguation is the problem of choosing the correct use of a word, given a set

Aplicaciones en CCSS

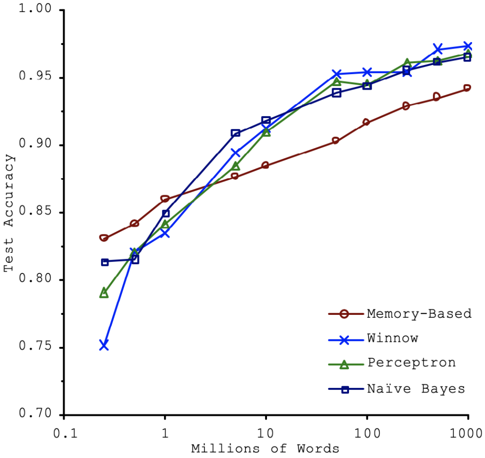


Figure 1. Learning Curves for Confusion Set Disambiguation

Aplicaciones en CCSS

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86. Association for Computational Linguistics.

Thumbs up? Sentiment Classification using Machine Learning Techniques

Bo Pang and **Lillian Lee**
Department of Computer Science
Cornell University
Ithaca, NY 14853 USA
{pabo,llee}@cs.cornell.edu

Shivakumar Vaithyanathan
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120 USA
shiv@almaden.ibm.com

Abstract

We consider the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, we find that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods we employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. We conclude by examining factors that make the sentiment classification problem more challenging.

use. Sentiment classification would also be helpful in business intelligence applications (e.g. MindfulEye's Lexant system[1]) and recommender systems (e.g., Terveen et al. (1997), Tatemura (2000)), where user input and feedback could be quickly summarized; indeed, in general, free-form survey responses given in natural language format could be processed using sentiment categorization. Moreover, there are also potential applications to message filtering; for example, one might be able to use sentiment information to recognize and discard "flames" (Spertus, 1997).

In this paper, we examine the effectiveness of applying machine learning techniques to the sentiment classification problem. A challenging aspect of this problem that seems to distinguish it from traditional topic-based classification is that while topics are often identifiable by keywords alone, sentiment can be

Aplicaciones en CCSS

	Proposed word lists	Accuracy	Ties
Human 1	positive: *dazzling, brilliant, phenomenal, excellent, fantastic* negative: *suck, terrible, awful, unwatchable, hideous*	58%	75%
Human 2	positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting* negative: *bad, cliched, sucks, boring, stupid, slow*	64%	39%

Figure 1: Baseline results for human word lists. Data: 700 positive and 700 negative reviews.

	Proposed word lists	Accuracy	Ties
Human 3 + stats	positive: *love, wonderful, best, great, superb, still, beautiful* negative: *bad, worst, stupid, waste, boring, ?, !*	69%	16%

Figure 2: Results for baseline using introspection and simple statistics of the data (including *test* data).

Aplicaciones en CCSS

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Figure 3: Average three-fold cross-validation accuracies, in percent. Boldface: best performance for a given setting (row). Recall that our baseline results ranged from 50% to 69%.

Human Decisions and Machine Predictions
Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan
NBER Working Paper No. 23180
February 2017
JEL No. C01,C54,C55,D8,H0,K0

ABSTRACT

We examine how machine learning can be used to improve and understand human decision-making. In particular, we focus on a decision that has important policy consequences. Millions of times each year, judges must decide where defendants will await trial—at home or in jail. By law, this decision hinges on the judge's prediction of what the defendant would do if released. This is a promising machine learning application because it is a concrete prediction task for which there is a large volume of data available. Yet comparing the algorithm to the judge proves complicated. First, the data are themselves generated by prior judge decisions. We only observe crime outcomes for released defendants, not for those judges detained. This makes it hard to evaluate counterfactual decision rules based on algorithmic predictions. Second, judges may have a broader set of preferences than the single variable that the algorithm focuses on; for instance, judges may care about racial inequities or about specific crimes (such as violent crimes) rather than just overall crime risk. We deal with these problems using different econometric strategies, such as quasi-random assignment of cases to judges. Even accounting for these concerns, our results suggest potentially large welfare gains: a policy simulation shows crime can be reduced by up to 24.8% with no change in jailing rates, or jail populations can be reduced by 42.0%with no increase in crime rates. Moreover, we see reductions in all categories of crime, including violent ones. Importantly, such gains can be had while also significantly reducing the percentage of African-Americans and Hispanics in jail. We find similar results in a national dataset as well. In addition, by focusing the algorithm on predicting judges' decisions, rather than defendant behavior, we gain some insight into decision-making: a key problem appears to be that judges to respond to 'noise' as if it were signal. These results suggest that while machine learning can be valuable, realizing this value requires integrating these tools into an economic framework: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals.

Innovative Ideas and Gender Inequality

Marlène Koffi
University of Toronto *

January 7, 2021

Abstract

This paper analyzes the recognition of women's innovative ideas. Bibliometric data from research in economics are used to investigate gender biases in citation patterns. Based on deep learning and machine learning techniques, one can (1) establish the similarities between papers (2) build a link between articles by identifying the papers citing, cited and that should be cited. This study finds that, on average, a paper omits almost half of related prior papers. There are, however, substantial heterogeneities among the authors. In fact, omitted papers are 15% to 30% more likely to be female-authored than male-authored. First, the most likely to be omitted are papers written by women (solo, mostly female team) working at mid-tier institutions, publishing in non-top journals. In a group of related papers, the probability of omission of those papers increases by 6 percentage points compared to men in similar affiliation when the citing authors are only males. Overall, for similar papers, having at least one female author reduces the probability of omitting other women's papers by up to 10 percentage points, whereas having only male authors increases the probability of being omitted by almost 4 percentage points.

Aplicaciones en CCSS- Perú

Using Neural Networks to Predict Micro-Spatial Economic Growth
Arman Khachiyan, Anthony Thomas, Huye Zhou, Gordon H. Hanson, Alex Cloninger, Tajana Rosing, and Amit Khandelwal
NBER Working Paper No. 29569
December 2021
JEL No. R0

ABSTRACT

We apply deep learning to daytime satellite imagery to predict changes in income and population at high spatial resolution in US data. For grid cells with lateral dimensions of 1.2km and 2.4km (where the average US county has dimension of 55.6km), our model predictions achieve R2 values of 0.85 to 0.91 in levels, which far exceed the accuracy of existing models, and 0.32 to 0.46 in decadal changes, which have no counterpart in the literature and are 3-4 times larger than for commonly used nighttime lights. Our network has wide application for analyzing localized shocks.

Arman Khachiyan
Economics Department
UC San Diego
arman.khachiyan@gmail.com

Anthony Thomas
Department of Computer Science and Engineering
UC San Diego
San Diego, CA 92093
ahthomas@eng.ucsd.edu

Alex Cloninger
Department of Mathematics
San Diego, CA 92093
acloninger@ucsd.edu

Tajana Rosing
Department of Computer Science and Engineering
UC San Diego
San Diego, CA 92093
tajana@eng.ucsd.edu

Aplicaciones en CCSS

Journal Title
XX(X):1–11
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Using Regression Forest to Detect Corruption at the local level: Evidence from Perú

P. Coronado[1], C. Muñoz[2] and A. Quispe[3]

Abstract
This paper uses a novel dataset on corruption in Peruvian local governments to estimate a regression forest algorithm. The algorithm utilizes thousands of variables from Peruvian fiscal accounts, local government surveys, electoral characteristics, and many other sources to train a set of forest predictors. Among the estimated algorithms, the regression forest outperforms state-of-the-art gradient boost and random forest algorithms. The trained algorithm can be used by the government agency combating corruption in Peru to improve efficiency in the anti corruption policy.

Keywords
Regression Forest, Corruption, machine learning, random forest

Introduction

Corruption remains a significant obstacle in the complex interplay of governance systems, eroding public confidence and intensifying pre-existing inequalities Avis et al. (2018); Becker and Stigler (1974). In the administrative structures of Peruvian municipalities, systemic gaps significantly impact the provision and accessibility of public services. This burden falls disproportionately on the poor, who often allocate a larger portion of their income to bribes for essential services, including police and healthcare. On average, 5.11% of income is spent on bribes, underscoring the regressive nature of corruption and its profound impact on vulnerable populations. Ciudadana (2006); de la República (2016); Yamada and Montero (2011).

While traditional methods like audits and empirical inquiries provide valuable insights, they have intrinsic limitations. The emergence of machine learning in corruption detection represents a pivotal development, transitioning from traditional methods like audits and empirical investigations, which, despite their value, are limited by the multifaceted nature of corruption. This complexity, encompassing factors like individual behavior, procedural anomalies, and linguistic biases, necessitates more sophisticated approaches. Artificial intelligence, especially machine learning, has been increasingly acknowledged in academic research for its

instances of corruption. Utilizing comprehensive datasets, including fiscal records from Peru's Ministry of Economy and Finance and RENAMU survey data, the study leverages sophisticated information extraction methods (NLP) to build a detailed dataset. The resulting analysis, producing 'Political Process Indicators' and 'Corruption Dependent Variables', offers a groundbreaking framework that enhances understanding of municipal corruption's contributing factors and manifestations, significantly advancing the transparency and accountability in local governance.

In our study, we employ a regression forest classifier in corruption detection, achieving an accuracy of 87.2% and an F1-score of 66.6%. It integrates advanced analytics with extensive empirical data, aiming to enhance understanding of corruption in public governance. Set in an era of growing data and computational capabilities, this research pioneers data-driven methodologies for corruption detection and prevention, potentially revolutionizing transparency and accountability in local governance, particularly in Peru.

The rest of the paper is organized as follows. In Section 2 we present the Institutional Background and Related Works about Detecting Corruption. Section 3 describes the data sources used to build the corpus for corruption detection. Section 4 shows Machine Learning and Predictive Analysis in Detecting Corruption. Section 5 concludes.

Aplicaciones en CCSS- Perú

Inteligencia Artificial

Machine Learning- Definición

Aplicaciones en CCSS

Agricultural crop production estimation in northwestern Peru using convolutional neural networks and high-resolution remote sensing imagery

Pavel Coronado Castellanos[1*†] and Pedro Achanccaray Diaz[2†]

[1*]Artificial Intelligence and Computational Methods in Social Sciences Lab - QLAB, Pontifical Catholic University of Peru, Universitaria St. 1801, San Miguel, 15088, Lima, Peru.
[2]Institute of Geodesy and Photogrammetry - IGP, Technical University of Braunschweig, Bienroder Weg 81, Braunschweig, 38106, Lower Saxony, Germany.

*Corresponding author(s). E-mail(s): pavel.coronado@pucp.edu.pe ;
Contributing authors: p.diaz@tu-braunschweig.de;
†These authors contributed equally to this work.

Abstract

The Earth's population growth has continuously increased the demand for agricultural production. Consequently, acreage and crop yield information have become increasingly important. Techniques based on satellite images are one of the most attractive options for agricultural monitoring over large areas. In this context, this work proposes a methodology for estimating the production of agricultural crops. Our study area is the La Libertad department, located in the northwest of Peru. The proposed methodology is based on convolutional neural networks (CNN) trained on high-resolution remote sensing imagery. The CNN is trained in an active learning cycle to assist in the labelling of new areas, to increase the dataset, and to improve the model's performance. The dataset used to train the CNN models has been developed in cooperation with the Ministry of Agrarian Development and Irrigation (MIDAGRI), which conducted field visits and surveys on a statistically representative sample of plots. Our experiments demonstrated that the proposed methodology is able to recognize different crops, even with only a few labelled parcels. Finally, the trained model is used to increase the number of labelled parcels for further improvements.