Seminario
Cali - Junio 07 de 2019

Tarea 1

Kevin García[1,a], Alejandro Vargas[1,b]

[1]Departamento de estadística, Universidad del Valle, Cali, Colombia

1. Antecedentes trabajo de grado

Se presentó una gran dificultad en la búsqueda de antecedente recientes sobre nuestro trabajo de grado, siendo conscientes de que los artículos posteriormente presentados no son ideales para el proceso de replicación, creemos que son los que más se acercan al ideal y con los cuales se podría tratar de hacer un tipo de replicación.

1.1. Planes de muestreo de aceptación de atributos mejorados basados en muestreo de nominaciones máximas

Este artículo, de los autores Jozani & Mirkamali (2010), demuestra el uso de la técnica de muestreo de nominación máxima (MNS) en diseño y evaluación de planes de muestreo de aceptación únicos para atributos. Se estudian tres tipos de niveles críticos de calidad: AQL(Acceptable Quality Level), LTPD(Lot Tolerance Percent Defective) y EQL(equilibrium quality level). Se analiza el efecto del tamaño de la muestra y el número de aceptación en el rendimiento de los planes MNS propuestos mediante la curva característica de operación(OC). Entre otros resultados, se muestra que los planes de muestreo de aceptación de MNS con un tamaño de muestra más pequeño y un número de aceptación mayor se desempeñan mejor que los planes de muestreo de aceptación comúnmente usados para atributos basados en la técnica de muestreo aleatorio simple(SRS). De hecho, los planes de muestreo de aceptación MNS dan como resultado curvas de OC que, en comparación con sus contrapartes de SRS, están mucho más cerca de la curva de OC ideal.

Estos planes propuestos se pueden usar de manera eficiente para situaciones en que el manejo de la calidad de los artículos es difícil, destructiva o muy costosa,

[a]Universidad del Valle. E-mail: kevin.chica@correounivalle.edu.co
[b]Universidad del Valle. E-mail: jose.alejandro.vargas@correounivalle.edu.co

pero los elementos a muestrear pueden clasificarse entre sí (según el criterio subjetivo o alguna medida auxiliar) en función de su nivel de calidad. Este artículo es muy importante para nosotros por el hecho de que hace parte de nuestro trabajo de grado, omitiendo lo anterior, consideramos que debe ser presentado en el seminario, ya que, en las técnicas de control de calidad empleadas en la práctica, y en casi cualquier tipo de industria, las decisiones de aceptar o rechazar lotes normalmente se basan en muestras extraídas de diferentes lotes y en la construcción de un plan de muestreo de aceptación. Además, el uso de estas técnicas no se ha quedado solo en la industria, ésta se ha expandido al ámbito de la salud, logrando cosas importantes como detectar problemas de calidad en métodos o procesos hospitalarios y encontrando problemas de cobertura en programas de salud. Esperamos que con la presentación de este tema, los compañeros posteriormente egresados, tengan claro el uso de esta metodología y logren aplicarla en su vida laboral si es requerida.

1.2. Planes de muestreo para numero de aceptación cero

Cuando se habla de control estadístico de calidad siempre se ha aceptado que no existe producto perfecto y que el "defecto" casi siempre existe en los lotes producidos, lo que se quiere lograr a la larga es disminuir esta cantidad de "defectuoso" logrando asi cumplir con los estándares de calidad que tiene cada organización.

Actualmente las organizaciones no se están conformando con tener una cantidad de defectuoso mínima si no que por el contrario quieren que sus lotes no contengan ningún artículo o producto defectuoso esto lleva a que sean necesarias el uso de estrategias de muestreo que logren obtener una muestra 0 defectuosa con una confianza alta de que el lote este "bueno" (libre de productos defectuoso).

En nuestro caso particular tenemos lotes de plantas de cítricos que pueden estar infectas con enfermedades y no es posible aceptar lotes con una o mas plantas infectadas por lo que es necesario conocer y aplicar un plan de muestreo que involucre como condición principal el no aceptar lotes que en la muestra tengan plantas infectas, así pues, Nicholas L. Squeglia nos muestra en su libro varias alternativas que podemos utilizar; como cada situación es distinta y no existe una verdad absoluta en cuanto a que plan de muestreo usar en qué situación, Nicholas también compara estas alternativas para que nosotros optemos por la que mejor nos convenga.

2. Efron & Hastie (2017)

2.1. Redes neuronales y aprendizaje profundo(cap 18)

El aprendizaje profundo y las redes neuronales hacen parte de la inteligencia artificial o aprendizaje automatizado, esto es, que por medio de un conjunto de algoritmos se intente modelar abstracciones de alto nivel en datos usando arquitecturas compuestas de transformaciones no lineales múltiples.

Las redes neuronales como su nombre lo indica, trata de asemejarse a lo que vendría siendo un cerebro, conectando "neuronas artificiales" de manera que simulen el aprendizaje humano, como puede llegar a ser el reconocimiento de un rostro, la compresión de algún lenguaje o incluso el reconocimiento de voz. Esto es posible gracias al análisis de cantidades gigantescas de datos y en algunos casos, se debe realizar en procesadores de alto rendimiento.

Este tema es interesante para nosotros ya que además de hacer uso de la estadística para entrenar estas redes, las cosas que se puede llegar a realizar en diferentes áreas son muy interesantes y podrían facilitar y automatizar de gran manera muchos procesos utilizando la tecnología. Por ejemplo, en la medicina, se podría usar este método para el reconocimiento de enfermedades en la piel por medio de una app para teléfonos inteligentes o en la agricultura, con la recolección de información por medio de drones para la toma de decisiones.

2.2. Bosques aleatorios e impulso(cap 17)

El concepto de bosques aleatorios surge con los árboles de decisión o diagramas de decisión, un árbol de decisión es un "mapa" de los posibles resultados de una serie de decisiones relacionadas, esto permite que una organización o individuo comparen diferentes acciones con el fin de maximizar la probabilidad de llegar al resultado esperado, esto puede ser por ejemplo conocer cuál es la ruta más óptima para llegar al trabajo o cuál va a ser la campaña y/o producto que me va a generar la mayor utilidad.

Los árboles de decisión se llaman de esta manera ya que asemejan las ramificaciones de los arboles de forma que comienza con un único nodo y se va ramificando en torno a los resultados que arroja (decisiones que se toman), cada nodo tiene una cierta probabilidad de tomar x o y opción y de esta manera se va expandiendo hasta llegar a una única respuesta, es aquí cuando termina.

¿Qué sucede cuando tenemos gran cantidad de datos? Bueno, pasamos de tener un árbol de decisión a una infinidad de árboles, esto en función a la cantidad de datos y lo que queramos hacer con estos, así pues, pasamos a tener un bosque aleatorio. Así como lo son las redes neuronales, los bosques aleatorios logran tener excelentes resultados en cuanto al modelamiento y la inteligencia artificial.

Nos vimos interesados en este tema ya que además de ser útil y moderno, este método facilita en gran medida el proceso de toma de decisiones tanto en organizaciones como en proyectos personales, además, es muy utilizado en problemas de clasificación(clasificar en clientes buenos y malos, clasificar modelos de coches distintos, clasificación de muestras de ADN para determinar enfermedades,entre otros), arrojando muy buenos resultados.

3. Lantz (2013) o Torgo (2017)

3.1. Aprendizaje probabilístico - clasificación usando bayes ingenuo(cap 4 - Lantz 2013)

El aprendizaje probabilístico, expuesto por Lantz (2013) en el cuarto capitulo de su libro, es un tema que nos pareció muy interesante ya que utiliza la estadística bayesiana, más específicamente el teorema de Bayes, para tratar de predecir o clasificar, modelando la incertidumbre por medio de probabilidades y utilizando internamente pruebas de hipótesis.

Investigando un poco a fondo sobre este tema, nos dimos cuenta que es un método muy utilizado en la actualidad en diversos contextos por su uso práctico y los buenos resultados en la mayoría de los casos. Por ejemplo, se puede utilizar para algo muy trivial, como clasificar una persona en hombre o mujer basándonos en las características de sus medidas: peso, altura y número de pie; pero, también se puede usar para problemas más complejos y muy útiles como por ejemplo, el que plantea el autor, filtrar el spam de los teléfonos celulares para eliminarlos automáticamente y que esto no se convierta en una molestia para los usuarios. Este último ejemplo, nos llamó mucho la atención y de aquí que este tema ocupara la primera posición, ya que los llamados "spam" son en la mayoría de casos cadenas de texto haciendo publicidad o informándonos acerca de un "premio" que ganamos sin siquiera participar, entre otros. Con la proliferación de datos no estructurados, la clasificación de texto o la categorización de texto ha encontrado muchas aplicaciones en la clasificación de temas, análisis de sentimientos, identificación de autoría, detección de correo no deseado, etc.

A pesar de que actualmente hay muchos algoritmos de clasificación disponibles, el Bayes ingenuo sigue siendo uno de los clasificadores más antiguos y populares. Por un lado, la implementación del bayes ingenuo es simple y, por otro lado, requiere menos cantidades de datos de entrenamiento. Sin embargo, el Bayes ingenuo se desempeña pobremente en comparación con otros clasificadores en la clasificación de texto. Como resultado, esto hace que el clasificador bayesiano ingenuo sea inutilizable en algunos casos a pesar de la simplicidad e intuición del modelo. Pero creemos que sería muy útil tanto para nosotros como para el grupo en general, saber como funciona este método y ver los posibles usos que este tiene en la vida cotidiana, ya que esto funcionaria como introducción al tema de algoritmos clasificadores y posiblemente nos facilite comprender la lógica y el funcionamiento de los nuevos algoritmos de clasificación que existen en la actualidad.

3.2. Divide y vencerás - Clasificación usando reglas y arboles de decisión(cap 5 - Lantz 2013)

La clasificación usando reglas y arboles de decisión, expuesto por Lantz (2013) en el quinto capitulo de su libro, al igual que el tema anterior, nos pareció interesante ya que básicamente, es un sistema de aprendizaje supervisado que aplica la estrategia "divide y vencerás" para hacer la clasificación, implementando méto-

dos y técnicas para la realización de procesos inteligentes, representando así el conocimiento y el aprendizaje, con el propósito de automatizar tareas. Su utilización y metodología es algo semejante al tema anterior, ya que dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

"Por su estructura son fáciles de comprender y analizar; su utilización cotidiana se puede dar en diagnósticos médicos, predicciones meteorológicas, controles de calidad, y otros problemas que necesiten de análisis de datos y toma de decisiones", lo que quiere decir que los árboles de decisión pueden ser usados en cualquier ámbito sin importar que sea laboral o personal, siempre y cuando implique toma de decisiones con cierto grado de incertidumbre (Calancha Zuniga, Carrión Bárcena, Cori Vargas, & Villa Torres, 2010, pág. 2).

Es evidente que obtener un mecanismo como este sería muy útil debido a que podríamos predecir comportamientos futuros a partir de los comportamientos observados en el pasado. Por ejemplo, a partir de los síntomas que tenemos observados en enfermos anteriores, y sabiendo ya si han desarrollado o no cierta enfermedad, podríamos extraer patrones que nos permitieran predecir si un paciente nuevo, aquejado de ciertos síntomas, desarrollará o no la misma enfermedad, lo que nos permitiría adelantarnos a su tratamiento. Otro ejemplo que es muy usado en la actualidad se encuentra en el mundo de los créditos bancarios, a partir de los comportamientos de los clientes antiguos con respecto a la morosidad o no de sus pagos del crédito concedido, podemos inferir qué nuevos clientes pueden ser los más convenientes para la concesión de un crédito, es decir, cuáles de ellos tienen más probabilidad de hacer frente al pago del mismo y cuáles más probabilidad de dejarlo sin pagar.

Para nosotros sería importante presentar este tema en seminario, ya que es una técnica muy utilizada por sus buenos resultados, y con ella, se abre paso a otra técnica mucho más potente que son los bosques aleatorios, los cuales utilizan una serie de arboles de decisión con el fin de mejorar la tasa de clasificación; por lo tanto, exponer este tema y ver sus posibles aplicaciones puede ser muy útil para todos, ya que en la vida laboral nos podemos encontrar tranquilamente con problemas de este tipo, que deben ser solucionados utilizando esta metodología.

References

Efron, B. & Hastie, T. (2017), *Computer age statistical inference. Algorithms, Evidence, and Data Science*, Cambridge.

Jozani, M. J. & Mirkamali, S. J. (2010), 'Improved attribute acceptance sampling plans based on maxima nomination sampling', *Statistical Planning and Inference* .

Lantz, B. (2013), *Machine learning with R*, Packt Publishing.