Empleo de Machine Learning para la predicción de los sitios de localización celular de proteínas

Jean Pierre López Villafani Luis Giancarlo Sánchez Huamán

20160809 20161333

Universidad de Lima Universidad de Lima

**CONCEPTOS**

**Regresión Logística Multinomial:** Extensión del método de regresión logística binaria clásica que emplea una variable dependiente de tipo nominal que cuenta con dos o más categorías (politómica). Las variables independientes pueden ser continuas o categóricas (Fernández, 2004). Se emplea para predecir las probabilidades de los diferentes resultados a obtener.

**Red Neuronal del tipo MLP:** Tipo de red neuronal, considerado por muchos como la más común y fácil de usar, se basa en el empleo de la red perceptrón simple empleando un número de capas ocultas, ya que se consideró que está ultima tenía varias limitaciones. Empleado para aproximar relaciones no lineales entre datos de entrada y salida.

**Secuencia de proteínas (protein sequencing):** Proceso para determinar la secuencia de aminoácidos de una parte o de toda una proteína.

**Keywords**

Secuencia de aminoácidos, protein sequencing, Localización celular de proteínas, Machine Learning, biología computacional, red neuronal del tipo MLP, regresión logística multinomial

# INTRODUCCIÓN

El campo de la biología, específicamente la biología computacional, está en constante crecimiento gracias a las nuevas tecnologías que están surgiendo. Un ejemplo de esto es el empleo de algoritmos con base de datos para predecir o determinar ciertas situaciones anómalas que se presenten dentro del cuerpo humano, con el fin de poder obtener un diagnostico antes de que la situación se pueda tornar irreversible. El problema radica en la dificultad para conseguir precisión en la respuesta, ya que los modelos que se diseñan pueden trabajar con cierta incertidumbre que puede generar problemas para el paciente, ya que se le puede dar un diagnóstico erróneo. Por tal motivo, este campo de estudio aún está en crecimiento; creando, aceptando y evaluando propuestas, con el propósito de obtener un modelo idea. Actualmente no existe una forma específica de tratar la data para poder obtener el mejor rendimiento posible, sino existen varios métodos que se ajustan a lo que el investigador está desarrollando. Por tal motivo, el empleo de esta metodología se utiliza para la comparación de resultados de diferentes modelos que existen.

Actualmente la aplicación de Machine Learning en el campo de la biología está teniendo una gran acogida, ya que se está realizando varias investigaciones respecto a este ámbito que proporciona una gran cantidad de información para que se pueda predecir el resultado, por ejemplo, en la rama de la biología molecular y celular de genomas humanos.

El proyecto por desarrollar en este artículo consiste en predecir los sitios de localización de secuencias de proteínas dando distintos datos, los cuales serán explicados en puntos posteriores. El propósito de conocer sobre la localización celular de la proteína consiste en poder aclarar su función en los healthy processes, así como su uso en fármacos que ayuden contra la aparición de enfermedades (Chen, 2008). El trabajo que se lleva actualmente consiste en emplear la caracterización experimental, la cual es precisa pero lenta y laboriosa. Estudios científicos revelaron una relación crucial entre la secuencia de aminoácidos de una proteína con el sitio de localización celular. Para llevar a cabo esta experimentación, empleando métodos computacionales como Machine Learning, se toma en cuenta diferentes métodos de localización, como de reconocimiento de señales (McGeoch o Von Heijne) entre otros. El problema de emplear solo uno de estos métodos es que determina si pertenece a una categoría de proteína por medio de una probabilidad; es decir, el modelo te indica si es, por ejemplo, una proteína de núcleo o no. Por tal motivo, se tiene que emplear varios métodos para determinar a qué categoría pertenece el dato seleccionado. El proceso se debe realizar de un dato (secuencia de aminoácidos) frente a todos estos métodos, ya que es necesario que se compare todos los resultados para determinar cuál es el más confiable y tener la certeza a cuál no pertenece. En conclusión, el empleo de técnicas de Machine Learning ayudará a que se pueda determinar de una manera mucho más rápida el análisis de la secuencia de aminoácidos (protein sequence), permitiendo mayores facilidades para el estudio en este campo de la biología.

En este artículo emplearemos dos métodos de Machine Learning, regresión logística de tipo multinomial y una red neuronal de tipo MLP, para poder realizar una comparativa de resultados y lógica, al mismo tiempo que evaluamos los resultados obtenidos al trabajar con un dataset de localizador celular de proteínas, el cual cuenta con información relacionada a los diferentes métodos de señal de secuencia de aminoácidos.

# INFORMACIÓN EMPLEADA

Para poder llevar a cabo esta investigación, era necesario contar con información relacionada al estudio de protein sequencing, por tal motivo se decidió emplear el data set del investigador Kenta Nakai, del Instituto de Biología Molecular de la Universidad de Osaka, el cual fue obtenido en la página web “<http://archive.ics.uci.edu/ml/datasets/yeast>” de la Universidad de California en Irvine (UCI), específicamente en su repositorio dedicado a estudios de Machine Learning.

Este dataset cuenta 1484 instancias, 9 atributos y no cuenta con valores en blanco, por tal motivo no se tiene que emplear técnicas para poder depurarlo. Además, la data con la que se trabaja es numérica, por tal motivo no se tiene que recurrir a técnicas de transformación de datos categóricos como ordinal encoding o one hot encoding. Por último, la data trabaja con un solo rango en sus datos, por eso no es necesario llevar a cabo una normalización.

A continuación, presentamos los atributos que se pueden encontrar dentro del data set Yeast Data Set.

* Nombre de la Secuencia: Representa la etiqueta correspondiente a la base de datos SWISS-PROT.
* Mcg: Representa el valor obtenido en el método McGeoch para el reconocimiento de la señal de secuencia.
* Gvh: Representa el valor obtenido en el método von Heijne para el reconocimiento de la señal de secuencia.
* Alm: Representa el valor de la región de expansión de la membrana ALOM.
* Mit: Representa el análisis discriminante del contenido de aminoácidos de la región N-terminal (20 residuos de largo) de proteínas mitocondriales y no mitocondriales.
* Erl: Representa la subcadena "HDEL" (se cree que actúa como una señal de retención en la luz del retículo endoplásmico). Es un atributo binario.
* Pox: Representa la señal de direccionamiento peroxisomal en el C-terminal.
* Vac: Representa el valor del análisis discriminante del contenido de aminoácidos de las proteínas vacuolares y extracelulares.
* Nuc: Representa el valor del análisis discriminante de las señales de localización nuclear de proteínas nucleares y no nucleares.

Adicional se cuenta con otra columna, que contiene información relevante para llevar a cabo un modelo supervisado, el cual decidimos llamar “rpta”.

* rpta: Representa la respuesta que debería obtener el modelo tras analizar la data presentada anteriormente.

Dentro de la clase respuesta podemos encontrar los siguientes datos cytosolic or cytoskeletal (CYT), nuclear (NUC), mitochondrial (MIT), membrane protein without Nterminal signal (ME3), membrane protein with uncleaved signal (ME2), membrane protein with cleaved signal (ME1), extracellular (EXC), vacuolar (VAC), peroxisomal (POX), endoplasmic reticulum lumen (ERL) (Chen, 2008).

# EXPERIMIENTACIÓN

Para la elaboración de la experimentación se usó como lenguaje de programación “Python” empleando el IDE “Spyder 3.7”, y el data set mencionado anteriormente.

Antes de llevar a cabo los métodos de regresión lineal multinomial y red neuronal de tipo MLP, tenemos que examinar la data a emplear para el trabajo. Para ello tenemos que ver la relevancia que tiene el atributo, según la descripción descrita en el punto anterior, en la determinación de la localización de la célula. Tras analizar la descripción de cada atributo, identificamos que se tiene una etiqueta de la base de datos, la cual tiene nula relevancia para el análisis a llevar a cabo, por ellos procederemos a retirarlo. El tipo de aprendizaje empleado en ambos métodos es el Supervised Machine Learning.

Otro aspecto a tener en cuenta es que se esta trabajando con un modelo supervisado, por ende, debemos separar la respuesta de los atributos, para ellos lo separamos en dos variables “x” e “y”; y también, se tuvo que realizar la separación en grupos de data de entrenamiento y data de prueba. Para ello decidimos trabajar con 07 y 0.3 respectivamente para ambos métodos.

Ambos modelos tienen que leer el archivo “yeast.data”, para ello se empleó la librería pandas.

## Regresión Logística del tipo multinomial

Para realizar la prueba, se usó principalmente las librerías sklearn.linear\_model importando LogisticRegression, sklearn.metrics importando confusion\_matrix y sklearn importando model\_selection, además del uso de un archivo yeast.data con los datos a evaluar.

Con la data dividida en dos grupos, entrenamiento y prueba, se aplicó la función LogisticRegression(), en el que se realizó la prueba con los siguientes parámetros: max\_iter = 3000, multi\_class=”multinomial”, solver=”sag”. En estos parámetros se indica la cantidad de iteraciones a realizar, la clase a usar(“multinomial”), y el algoritmo a emplear para resolver el problema de optimización, en nuestro caso “sag”. La cantidad de iteraciones se tomó de manera arbitraria, y finalmente se entrenó al modelo dando los resultados que se mostrarán en el siguiente punto.

## Red neuronal del tipo MLP

Para realizar la prueba se empleó principalmente las siguientes librerias. klearn.neural\_network importante MLPClassifier, para poder clasficar el conjunto de datos; sklearn.model\_selection importando train\_test\_split, para dividir la data de entrenamiento de la de prueba, sklearn.metrics importando confusion\_matrix y matplotlib.pyplot para mostrar gráficamente los resultados.

Con la data dividida en dos grupos, entrenamiento y prueba, se aplica la función MLPClassifier() con los siguientes parámetros activation = ‘relu’, para activar la función de capa oculta e indicar que será una función de unidad lineal rectificada, batch-size = 4 para determinar el tamaño de minibatch a trabajar, el learning rate = “constant”, que es la tasa de aprendizaje inicial para controlar el tamaño de los pesos en cada actualización, max\_iter = 3000 para determinar las epocas a trabajar(en este caso se escogió de forma arbitraria) y el solber = “adam”, como el optimizador estocástico. Tras especificar el modelo de red neuronal a trabajar procederemos a “entrenarlo”

# RESULTADOS

Tras realizar ambos procedimientos procederemos a emplear en cada proyecto las librerías MLPClassifier, para el caso de la red neuronal MLP, LogisticRegresion, para la regresión lineal multinomial, confusión\_matrix y matplotlib.pyplot. Para poder realizar un análisis a los datos obtenidos en la experimentación.

Antes de mostrar los resultados obtenidos, procederemos a hacer un análisis a la correlación de los atributos dados, para ver si se presenta una alta o baja correlación entre ellas.

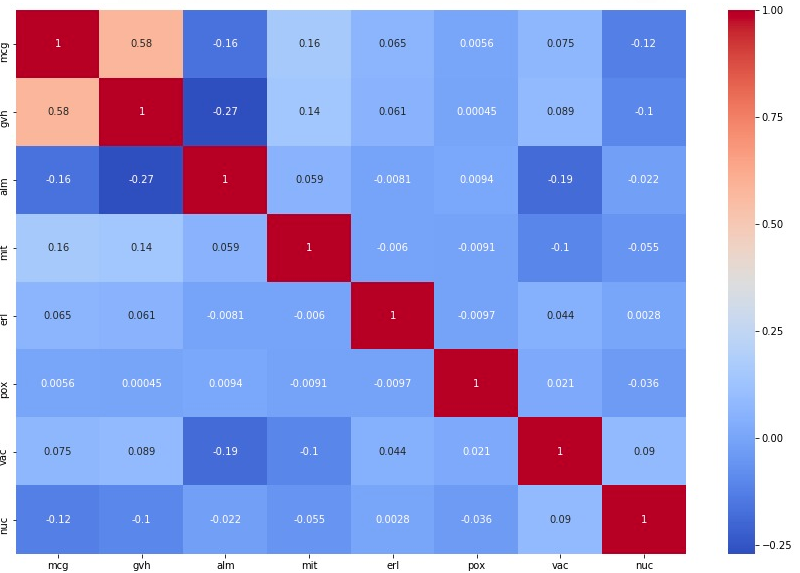


Figure 1. Matriz de correlación de atributos.

Gracias a este análisis podemos ver que existe una baja correlación entre la mayoría de los atributos, permitiéndonos comprobar que cada variable tiene influencia dentro del resultado a obtener (la localización celular de la proteína). En el caso de los atributos “mcg” y “gvh”, ambos tienen una correlación de 0.58 debido a que ambos evalúan el reconocimiento de secuencia de señal, pero empleando una metodología distinta, por tal motivo, se intuye que pueda existir una correlación entre estos atributos.

Para el caso de la red neuronal MLP se obtuvo la siguiente matriz de confusión con un porcentaje de precisión del 0.61 y un error-rate de 0.39 aproximado:

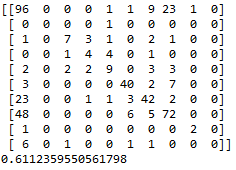


Figure 2. Matriz de confusión de la red neuronal MLP

Se puede observar que el porcentaje de precisión es aceptable para el modelo; sin embargo, el modelo muestra más errores de confusión en la primera columna.

Para el caso de la regresión logística multinomial se obtuvo la siguiente matriz de confusión con un porcentaje de precisión de 0.56 y un error-rate de 0.44 aproximado:

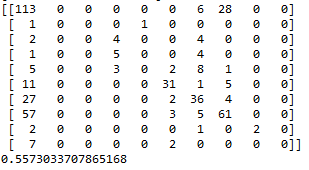


Figure 3. Matriz de confusión de la regresión logística multinomial

A continuación, presentamos una tabla que nos muestra los resultados obtenidos:

Table 1. Resumen de resultados de precisión de los modelos

|  |  |
| --- | --- |
| Precisión (Accuracy) | |
| Red neuronal MLP | Regresión logística multinomial |
| 0.61 | 0.56 |

Presentándose el mismo problema que el anterior, en el que la predicción no termina acertando con el valor indicado.

Adicionalmente, se realizó la curva ROC para la red neuronal de tipo MLP, con fin de poder ver el rendimiento del modelo en las épocas trabajadas, como se puede apreciar mientras más iteraciones se realicen(épocas) el modelo va reduciendo la tasa de Falsos Positivos, lo cual se debe mantener controlado para evitar que se presente un overfitting.

Imagen que contiene captura de pantalla, texto

Descripción generada automáticamente

Figure 4. Curva ROC de la red neuronal MLP.

En el caso de la Regresión Lineal Muntinomial el trazado de la curva no se pudo realizar debido a que se tenía que hacer uso de la metodología One vs All, debido a las múltiples clases que tenía el modelo. De esta forma se tendría que graficar más de una línea, teniendo que ser una línea por cada clase comparando su rendimiento con la suma de las otras.

# CONCLUSIONES

Las conclusiones obtenidas tras la elaboración de este trabajo son:

* El empleo de esta propuesta para determinar la localización celular de proteínas aún necesita mejorarse, ya sea ingresando una mayor cantidad de datos o buscando un modelo que se adapte mejor a la estructura trabajada.
* Al ser un escenario diseñado para esta experimentación se tuvo mayores facilidades de evaluación, ya que no se tuvo en consideración valores nulos o indefinidos.
* El modelo que obtuvo, ligeramente, mayor precisión fue el de la red neuronal MLP sobre la regresión logística multinomial, aunque si se realizan más corridas este número puede igualarse.
* No existe mucha diferenciación en el resultado de estos modelos, sería conveniente probar el mismo data set frente a otros modelos para evaluar los resultados obtenidos.
* La aplicación de ambos modelos en el lenguaje empleado no presenta demasiada dificultad al contar con librerías que apoyan a la implementación de estos.

# REFERENCIAS

Chen, Y. (2008). *Predicting the Cellular Localization Sites of Proteins Using Decision Tree.* Iowa,USA: Iowa State University .

Fernández, V. P. (2004). *Regresión Logística Multinomial.* Madrid,España: Cuadernos de la Sociedad Española de Ciencias Forestales.

Horton, P., & Nakai, K. (1996). *A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins.* St. Louis, USA: Intelligent Systems in Molecular Biology-96.