CIENCIA DE DATOS

Evaluación del desempeño de modelos de clasificación tras aplicar métodos de reducción de dimensionalidad y selección de características

Fernanda G. Brenda S. Alejandra S. Elizabeth S.
Clase de Ciencia de Datos 2020
Módulo: Aprendizaje de Máquina

Abstract

En este trabajo se usaron los métodos de reducción de dimensionalidad y selección de características para evaluar su impacto en la eficiencia de clasificación. Se obtuvo que en la mayoría de los casos, la reducción del tamaño de los datos aumenta el F1-Score de la clase positiva. Las características seleccionadas por diferentes métodos fueron similares, y la reducción de dimensionalidad permitió diferenciar las clases por medio de tSNE.

Introducción

En los métodos de clasificación, la alta dimensionalidad de los datos procesados es una limitante a la eficiencia de los clasificadores. Algunas consecuencias de esta problemática son la dificultad para visualizar los datos y el aumento en el tiempo de procesamiento. (1)

El objetivo de este trabajo es evaluar la eficiencia de distintos métodos de reducción de dimensionalidad en la predicción de dos conjuntos de datos, además de evaluaciones de parámetros propios de cada método para mejorar la clasificación y la relación entre la cantidad de datos y la eficiencia. Adicional a esto, se compararon las características seleccionadas de un conjunto de datos.

Usamos dos conjuntos de datos, el primero contiene oraciones de dominios estructurales proteicos, y el segundo tiene información de compuestos químicos y su habilidad de unirse a la trombina.

El conjunto de datos de oraciones de dominios tuvo mejor rendimiento en la clasificación que el conjunto de unión a trombina, y las características seleccionadas por distintos modelos fueron muy similares. Para la mayoría de los modelos del conjunto de unión a trombina, conforme aumenta la cantidad de características seleccionadas el rendimiento disminuye, aunque hay excepciones.

Materiales y métodos

Datos

Los datos de dominio estructural se obtuvieron del repositorio de Carlos Francisco Méndez Cruz, llamado "material suplementario para el aprendizaje supervisado el tema de clasificación". El conjunto consta de oraciones y se clasifica en dos clases de moléculas, aquellas que presentan un dominio y aquellas que no. El conjunto consiste de 2,237 datos, los cuales se dividieron en 1566 datos de entrenamiento y 671 datos de evaluación.

Los datos de trombina fueron obtenidos de la Copa KDD 2001. Este conjunto se clasifica en dos clases distintas de medicamentos, los que se unen y los que no se unen a la trombina. Se cuentan con 2,545 datos, los cuales se dividieron en 1,909 datos de entrenamiento y 634 datos de evaluación.

Métodos

Se usarán métodos de reducción de dimensionalidad y selección de características. La reducción de dimensionalidad, es la transformación o proyección de los datos en un espacio de menor magnitud. (2) Dentro de este grupo de métodos, usaremos Análisis de Componentes Principales (PCA), Descomposición en Valores Singulares (SVD) e Incrustación Estocástica de Vecinos t-distribuida (t-SNE).

Los métodos de selección de características que usaremos son Información Mutua (MI) y Chi2. Estos métodos se basan en seleccionar del grupo de datos las variables que se consideran más relevantes para la clasificación. Una de las ventajas de este grupo es que al no hacer transformaciones con los datos, son interpretados con mayor facilidad. (2)

Metodología de datos de dominio estructural
Para este conjunto, se llevaron a cabo 64 simulaciones, utilizando dos clasificadores distintos: Bayesiano Ingenuo Multinomial (MNB) y Máquina de Soporte Vectorial (SVM), y dos métodos de selección de características: MI y Chi2. Se probaron ambos clasificadores con los dos tipos de métodos de selección de características, y con 8 números distintos de características seleccionadas (k), los cuales fueron k=1, k=10, k=50, k=100, k=500, k=1,000, k=2,000 y k=3,000.

Asimismo, para la SVM, se probaron dos kernels distintos, uno lineal y uno RBF. Por otro lado, debido a la forma de operación de los clasificadores bayesianos, es necesario hacer una conversión de los datos para que éstos fueran positivos. Para esto, se probaron dos métodos distintos, aplicar valor absoluto y obtener los cuadrados de cada observación. Sin embargo, creemos que esto no es necesario para la selección de características debido a que los datos se encuentran como una matriz binaria y no se realiza una transformación de los datos, únicamente se seleccionan los que contengan mayor información. Realizamos estos dos métodos exclusivamente con la finalidad de confirmar dicha teoría.

Dado que el objetivo principal de este trabajo es tener una mejor clasificación de la clase positiva, se utilizó el F1-Score de la clase DOM, que corresponde a aquellas oraciones que presentan un dominio estructural para discriminar el rendimiento entre modelos de clasificación. Para una visualización gráfica de todas las combinaciones generadas, véase la figura suplementaria S1.

Finalmente, seleccionamos los 3 mejores modelos de clasificación tras el uso de MI y CHI2, y obtuvimos las características, es decir, las palabras seleccionadas por cada uno de ellos para compararlas.

Metodología de datos de unión a trombina
Para este conjunto, se llevaron a cabo 106 simulaciones. Se usaron los clasificadores MNB y SVM, dos métodos de selección de características: MI y Chi2, y tres métodos de reducción de dimensionalidad: PCA, SVD y tSNE.

MI y Chi2 se probaron con ambos clasificadores, utilizando 7 números distintos para k, los cuales fueron k=10, k=100, k=200, k=400, k=1,000, k=10,000 y k=100,000. Debido a que el conjunto de datos de dominio estructural mostró que no es necesario realizar una conversión hacia valores positivos con métodos de selección de características en un clasificador bayesiano, para este conjunto no se realizó este paso. La SVM nuevamente se probó con kernel lineal y RBF.

PCA y SVD se probaron con ambos clasificadores, utilizando 5 números distintos de componentes principales (PCs), los cuales fueron: 10 PCs, 100 PCs, 200 PCs, 400 PCs y 1,000 PCs. Estos componentes principales representan el 28.76%, el 63.82%, el 77.81%, el 90.01% y el 99.48% de la varianza, respectivamente. Para estos métodos, debido a que se realiza una transformación de los datos, se probaron los métodos de obtener el valor absoluto y el cuadrado de cada observación para entrenar al MNB. Por otro lado, la SVM se probó con kernel lineal y RBF.

Finalmente, tSNE se probó con ambos clasificadores, utilizando 3 números distintos de PCs: 1 PC, 2 PCs y 3 PCs. Asimismo, se utilizaron 3 valores distintos de perplejidad, los cuales fueron 10, 30 y 50. Al igual que para PCA y SVD, se probaron los métodos de valor absoluto y cuadrados de cada observación para la obtención de valores positivos para entrenar al MNB, pues se llevó a cabo una transformación de los datos iniciales. La SVM se probó con kernel lineal y RBF.

De igual forma que para dominio estructural, se utilizó F1-Score de la clase positiva, que en este caso es A (Activated) para discriminar la eficiencia de los modelos. Para una visualización gráfica de todas las combinaciones generadas, véase la figura suplementaria S2 y S3.

Por último, se visualizaron los resultados de la reducción de dimensionalidad por tSNE para el conjunto de datos iniciales y para los clasificadores con mejor puntuación correspondientes a los métodos de selección de características y de reducción de dimensionalidad. Estos clasificadores son Chi2 con 10 características y PCA con 1000 PCs, respectivamente. Asimismo, se utilizó el valor de perplejidad correspondiente al clasificador con tSNE cuya puntuación corresponde a la más alta, el cual fue 30.

Resultados

Dominio estructural
Los resultados de las iteraciones de este conjunto de datos se pueden observar en la figura 1 y la tabla suplementaria S1.

Podemos ver como en el clasificador de MNB (Figura 1A) tenemos mejores resultados que en SVM (Figura 1B), esto principalmente al momento de seleccionar más de 50 características.

Como habíamos intuido previamente, el tipo de conversión a datos positivos no afecta el rendimiento del clasificador MNB, pues se trata de una matriz binaria. Para este clasificador, resalta que tiene una alta tasa de crecimiento pasando de 0 con 1 característica seleccionada a 0.8 con 100 características (2.85% de datos totales), siendo este el F1-Score más alto. Posterior a esto alcanza un estado de equilibrio donde no

varía el rendimiento. Por otro lado, el clasificador SVM muestra mayor variación entre el rendimiento al usar diferentes kernel, y a diferencia de MNB, no tiene una alta tasa de crecimiento, sino que con la utilización de únicamente una característica tiene un rendimiento de 0.61.

Con respecto a los métodos de selección de características, los valores son muy similares. Globalmente, el mayor F1-Score es 0.80 para ambos métodos (MI y Chi2). Dicho score es alcanzado por MI con 100 características y por Chi2 con 500, usando MNB como clasificador en ambos casos. Para SVM, el F1-Score más alto fue 0.79, el cual fue alcanzado por Chi2 al utilizar 3,000 características. Seguido de valores de 0.78 por MI y Chi2, seleccionando 100 y 50 características, respectivamente. En general, vemos una mayor variabilidad en el rendimiento de los modelos entre las 50 y 1000 características.

Las características de mayor relevancia seleccionadas por los cinco mejores clasificadores mencionados en el párrafo anterior son palabras que tienen relación con el tema de proteínas: domain, terminal, helix, turn (Tabla suplementaria S2). Creemos que estas palabras representan correctamente la clase positiva del conjunto de datos de entrenamiento. "terminal" se refiere a los dominios N y C terminales a los extremos de las proteínas, y "helix" y "turn" corresponden a un tipo de dominio llamado "helix-turn-helix", presente en proteínas cuya función es de regulación transcripcional. (4)

Unión a trombina

Los resultados de las iteraciones de este conjunto de datos se pueden observar en la figura 2 y la tabla suplementaria S3.

Lo primero que podemos observar, es que para los métodos de selección de características, en general, los valores de F1-Score disminuyen conforme aumenta el número de características seleccionadas tanto con el clasificador SVM (Figura 2A), como con el MNB (Figura 2B).

Chi2 tiene el valor más alto de F1-Score en ambos clasificadores seleccionando 10 características y en el caso de SVM, el valor más alto fue alcanzado utilizando un kernel lineal (0.5 y 0.45, respectivamente). Posteriormente, observamos que en general, el F1-Score muestra una decaída conforme aumentan el número de características. Curiosamente, podemos notar que en la SVM con 400 características nuevamente suben los valores con ambos kernels, y con 100,000 únicamente con kernel RBF. En el caso de MI, los valores máximos se alcanzan al utilizar 100 características en ambos clasificadores, y en la SVM, nuevamente con un kernel lineal (0.34 y 0.38, respectivamente). Después de este punto, conforme k aumenta, notamos que la F1-Score disminuye gradualmente en ambos clasificadores, y en SVM con ambos kernels.

En cuanto a la reducción de dimensionalidad, observamos un decremento en F1-Score en todos los clasificadores SVM conforme aumentamos la cantidad de componentes (Figura 2D, F). De forma inversa, en todos los clasificadores de MNB la F1-Score aumenta conforme aumenta el número de componentes principales (Figura 2C, E). De forma general, los modelos parecen agruparse de acuerdo al tipo de kernel usado en el clasificador SVM.

El el caso de PCA, podemos observar que sus valores máximos se alcanzan al seleccionar 1,000 componentes principales con un clasificador MNB y obteniendo los cuadrados y 10 componentes principales con una SVM con la utilización de un kernel lineal (0.37 y 0.35, respectivamente). Sin embargo, con un kernel RBF, la puntuación se mantuvo en 0 con todos los componentes principales. Por otro lado, observamos que SVD alcanza sus valores máximos con un clasificador MNB al seleccionar 10 componentes y obtener su valor absoluto, y con una SVM al seleccionar 1,000 componentes con un kernel lineal, siendo éstos valores 0.37 y 0.34, respectivamente. Finalmente, para tSNE con un clasificador SVM, los valores son considerablemente bajos, lo que nos dice que el método no fue muy bueno. Sin embargo, con el clasificador de MNB, mejora considerablemente con respecto al número de componentes seleccionados. Los valores más altos se logran con tres componentes, alcanzando un máximo de 0.37 al utilizar perplejidad de 30. Además, vemos un comportamiento similar en los modelos cuando agrupamos por valor de perplejidad en el clasificador MNB, ya que se observan cambios en las tasas de crecimiento similares. Por otro lado, el tipo de conversión a datos positivos no parece tener influencia en el rendimiento. Una particularidad de tSNE radica en su escala, ya que las escalas de MNB y SVM de este método son las más diferentes.

De todos los métodos, el que obtuvo la mejor puntuación en todos los clasificadores corresponde a Chi2, y posteriormente, las mejores puntuaciones corresponden a MI con clasificador SVM y kernel lineal al seleccionar 100 características, PCA con clasificador MNB y método de cuadrados al seleccionar 1,000 componentes y a tSNE con clasificador MNB, método de valor absoluto y perplejidad de 30 al seleccionar 3 componentes.

La iteración de tSNE con mejores resultados fue al seleccionar 3 componentes y perplejidad de 30, por lo que graficamos estos resultados en 2D y 3D (Figura 3A y HTML suplementario Figura 1, respectivamente). En esta gráfica, los datos pertenecientes a la clase negativa se agrupan en una esfera de diámetro

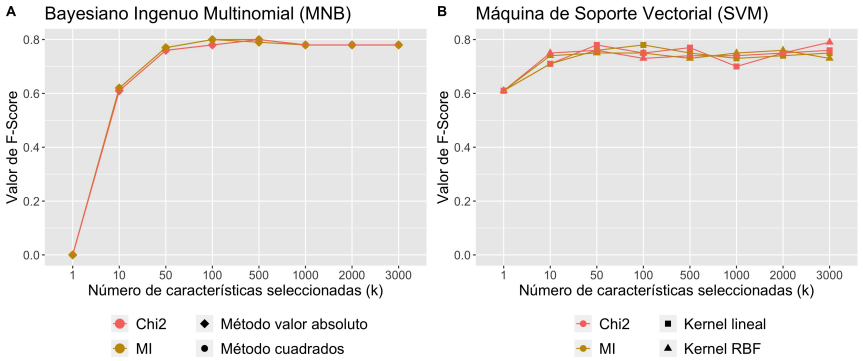

Figura 1: Resultados de los datos de dominio estructural con métodos de selección de características Chi2 y MI. A) Clasificador MNB. B) Clasificador SVM.

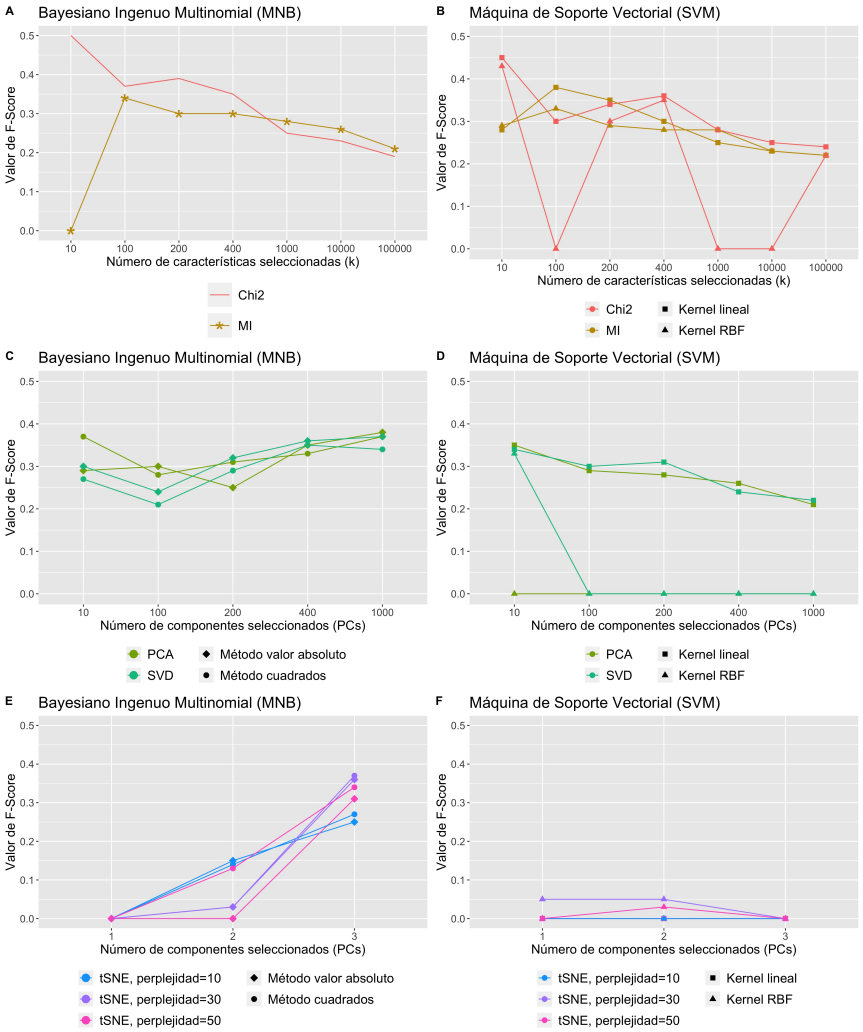

Figura 2: Gráficas de los resultados obtenidos con los datos de unión a trombina. Se encuentran del lado izquierdo los clasificadores MNB y del lado derecho las SVM. A) y B) Método de selección de características MI y Chi2. C) y D) Métodos de reducción de dimensionalidad PCA y SVD. E) y F) Métodos de reducción de dimensionalidad tSNE.

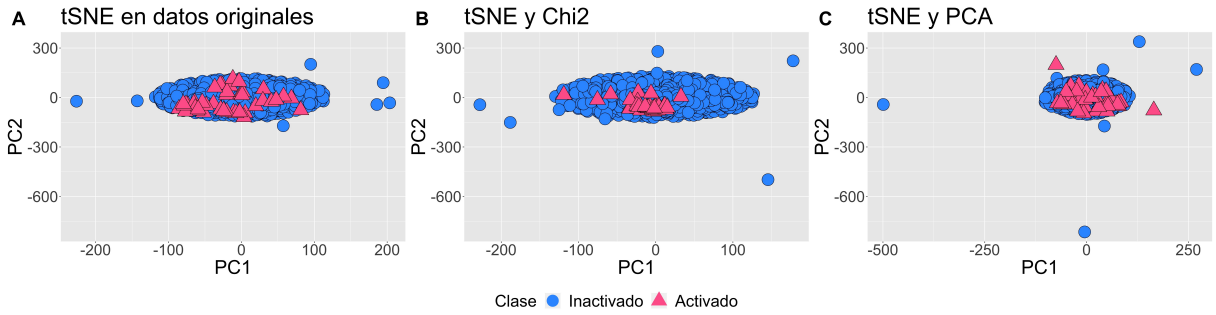

Figura 3: Gráficas de los mejores resultados con el método de tSNE. A)Datos originales B) Mejor método de características, Chi2. C) Mejor método de reducción de dimensionalidad, PCA.

cercano a 100. Mientras que los datos de la clase positiva se muestran más aplanados, de dimensiones cercanas a 80 x 70 x 50. Cabe mencionar que todos los datos de la clase positiva caen dentro del espacio de la clase negativa, por lo que la separación entre clases no es exitosa y parece restringirse a un solo eje del plano.

Posteriormente, escogimos los mejores resultados para selección de características y reducción de dimensionalidad para aplicarles tSNE con los parámetros que se aplicaron con anterioridad a los datos originales. El mejor método de selección de características fue Chi2 con 10 características y el mejor método de reducción de dimensionalidad fue PCA con 1000 características.

En el caso de PCA no observamos una separación en ninguno de los planos, no hay diferencias notables en comparación con tSNE aplicado a los datos originales (Figura 3B y HTML suplementario Figura 2). En el caso de Chi2, a pesar de que no hay una separación de las clases, podemos observar que la agrupación de las mismas ha mejorado, reduciendo el diámetro de la esfera (Figura 3C y HTML suplementario Figura 3). Este último punto es importante ya que el mejor F-Score obtenido en los clasificadores corresponde a las características seleccionadas por Chi2 con 10 elementos (F1-Score=0.5), en comparación con el de PCA con 1000 elementos que fue 0.38.

Conclusión

Estos resultados nos muestran que, en un mismo conjunto de datos el método con mejores resultados puede variar dependiendo del tipo de algoritmo de entrenamiento usado, los parámetros de éste, entre otros elementos variables. Sin embargo, no cabe duda de la utilidad de los métodos de selección de características y reducción de dimensionalidad para mejorar el rendimiento de los clasificadores; al mejorar la clasificación y al reducir el tiempo de procesamiento. Debido a esto, llegamos a la conclusión de que es bueno invertir en la exploración de estos métodos, siempre cuidando que no haya un sobre ajuste, de los datos.

Como perspectivas a futuro proponemos probar combinaciones de métodos de reducción de dimensionalidad y métodos de selección de características. Además, con estos datos sería bueno tratar desbalance de clases, ya que fue una problemática no atacada en esta investigación.

Referencias

1 Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. J Mach Learn Res, 10(66-71), 13.

2 Tran, B., Xue, B., & Zhang, M. (2014, December). Overview of particle swarm optimisation for feature selection in classification. In Asia-Pacific conference on simulated evolution and learning (pp. 605-617). Springer, Cham.

3 Stephen D. Bay and Dennis F. Kibler and Michael J. Pazzani and Padhraic Smyth. The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation. SIGKDD Explorations, 2. 2001

4 Dodd, I. B., & Egan, J. B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic acids research, 18(17), 5019-5026.

Archivos adicionales

Material suplementario

El material suplementario cuenta con graficas y tablas extras que permiten un mejor entendimiento de los métodos utilizados y los resultados obtenidos.

Repositorio

https://github.com/elisulvaran/StructuralDomain-BindingThrombin

Este repositorio cuenta con los códigos utilizados para correr los métodos, de igual forma cuenta con todos los archivos obtenidos y el archivo HTML con imágenes interactivas en 3D, de los modelos de la figura 3.