Feature Engineering

Al realizar procesos de reconocimiento de patrones ya sea mediante métodos supervisados, semi supervisados o no supervisados, uno de los elementos mas importantes a tener en cuenta son los descriptores o características que se usan para representar los datos. Estos descriptores son de suma importancia ya que la precisión del proceso depende de que éstos representen lo mejor posible los objetos a reconocer, además de que pueden ser también determinantes a la hora que los algoritmos usados sean eficientes en términos de recursos, tiempo, etc. Es por esto que existen diferentes tipos de técnicas usadas para procesar los datos crudos y obtener los datos que se usarán en el proceso de reconocimiento. Dependiendo del tipo de proceso realizado estas técnicas se pueden clasificar en tres grandes grupos Feature Construction, Feature Extraction y Feature Selection, de los cuales se pueden usar técnicas únicamente de uno de estos grupos o usar varias de ellas en conjunto según sea necesario en cada caso. Estos grupos se pasarán a describir a continuación.

Feature Construction: Este tipo de técnicas buscan generar nuevas características ampliando el espectro de datos del que se dispone inicialmente. Su uso principal es cuando se tienen datos incompletos y se desea buscar la información faltante que complemente la descripción del objeto o la relación que existe entre las características que ya se tienen. Esta nueva información se puede obtener por medio de inferencia o creación de nuevos datos, pero siempre usando los datos ya existentes para derivarla, de tal modo que no se agrega nueva información sino que se busca incrementar el poder de expresión de los datos ya existentes. Algunas de las técnicas que se podrían usar para generar estos nuevos datos se pueden categorizar en las dirigidas por datos, en las cuales se aplica algún operador sobre estos datos como por ejemplo operadores algebraicos en datos numéricos, operadores condicionales, conjunciones, disyunciones, etc.
Luego de realizar este proceso, muchos de estos nuevos datos generados pueden no ser relevantes por lo cual este tipo de técnicas suelen ir acompañadas con técnicas de Feature Selection para así descartar los datos redundantes o irrelevantes.

Feature Selection: En este tipo de técnicas, al contrario de las anteriores, se busca reducir el conjunto de datos a utilizar. Esta reducción puede ser realizada ya sea mediante una selección y descarte de las características disponibles, es decir, no existe ningún tipo de transformación u operación sobre los datos, o puede ser mediante agregación entre ellos. Este tipo de reducción o selección se puede realizar generando subconjuntos de formas, secuenciales, aleatorias, generar todos los posibles subconjuntos (2^N subconjuntos, donde N es el numero de características disponibles), etc. Dentro de los subconjuntos secuenciales podrían generarse de forma que se inicia con un subconjunto vacío y se van agregando características una a una (sequential forward selection) o empezar con el conjunto completo e ir eliminando características (sequential backward selection). Una técnica de selección aleatoria podría ser el Random Forest, en el cual se generan un conjunto de arboles de decisión, cada uno usando aleatoriamente un subconjunto de las características disponibles como nodos y dependiendo de los resultados arrojados en estos distintos arboles se seleccionan o descartan características, por ejemplo, si cada vez que estaba presente una característica **x** en un árbol de decisión era determinante en el resultado obtenido, entonces esa característica es de relevancia y debe ser seleccionada.
Se suelen usar diferentes criterios para evaluar si un subconjunto generado es óptimo o no, estos criterios se pueden agrupar dentro de dos grandes modelos: Filtering y Wrapping.

El **Filtering** consta de criterios que son independientes del modelo de aprendizaje que se usará luego. Es una aproximación eficiente pero debido a esta separación del proceso de aprendizaje posterior puede descartar información que si hubiera sido útil para el mejor rendimiento de ese algoritmo usado. El proceso usado suele realizarse en dos pasos; primero la aplicación del criterio de selección, entre las

cuales pueden estar: dependencia de una característica de su clase, correlaciones de característica-característica característica-clase, otras medidas de dependencia, medidas de distancia, medidas de información, medidas de consistencia, etc. En el segundo paso simplemente se selecciona el subconjunto con mejor ranking.

En el modelo **Wrapping** a diferencia de filtering sí se toma en cuenta el algoritmo de aprendizaje a usar, es decir, se mide el rendimiento obtenido por ese algoritmo usando el subconjunto de características seleccionado, este proceso se suele realizar usando una estrategia de búsqueda entre las características o subconjunto de características cuyo resultado se pasa al algoritmo se mide el rendimiento y este resultado se vuelve a pasar al componente de búsqueda en un proceso iterativo hasta que se selecciona el conjunto de características con el mejor rendimiento. Para evitar una búsqueda exhaustiva se pueden usar distintas estrategias como algoritmos genéticos, best-first, hill-climbing, etc.

Existe un tercer modelo usado llamado **Embedding** este modelo es una especie de unión entre filtering y wrapping para obtener lo mejor de ambos mundos. Ya que filtering no toma en cuenta el algoritmo de aprendizaje a usar pudiendo descartar características útiles para ese algoritmo y wrapping debe evaluar el rendimiento de los subconjuntos de característica en el modelo haciéndolo bastante costoso computacionalmente, el modelo embedding busca, como su nombre lo indica, incrustar el proceso de selección en la construcción del modelo de aprendizaje logrando mejorar el costo computacional ya que no es necesario correr el proceso de aprendizaje repetidas veces pero aún así tomando en cuenta la interacción con éste.

Feature Extraction estas técnicas buscan la transformación de los datos disponibles en datos que mejor representen y/o que mejor utilidad tengan para los modelos de aprendizaje que se vayan a usar posteriormente. Estas transformaciones suelen ser realizadas por medio de funciones de mapeo, para ilustrar esta idea se puede poner como ejemplo un caso donde como datos iniciales se tenga un conjunto de números de diferentes magnitudes (como 5, 1200, -50, 82154, etc.) a los cuales se les podría aplicar como función de mapeo la función sigmoide, acotando estos números a un espectro entre 0 y 1 que puede ser mas útil al momento de realizar la clasificación. Estas funciones de mapeo pueden variar dependiendo de distintos criterios como el tipo de característica sobre la cual se implementará, rendimiento que se busca, tipo de algoritmo de aprendizaje a usar, etc. aunque en los criterios mas comunes buscados están: maximizar varianza o variación, es decir, buscar características que tomen valores diferentes en cada instancia, ya que características que puedan tomar los mismos valores en diferentes instancias no aportarían ningún valor como discriminantes y reducir correlaciones o evitar características redundantes usualmente logrando esto por medio de funciones de mapeo que reducen la dimensionalidad.

Existen diferentes tipos de funciones que se pueden usar dependiendo del tipo de datos al cual se aplicará, por esto es importante determinar cuales son esos posibles tipos de datos:

Características estadísticas: Son características que se extraen de datos de tipo numéricos donde se busca representarlos mediante valores estadísticos como la media, mediana, moda, desviación estándar, etc. En este tipo de técnicas se puede incluir el Análisis de Componentes Principales (PCA por siglas en inglés) una de las mas comunes dentro del feature extraction en la cual se buscar reducir dimensionalidad buscando proyectar valores en componentes que mejor reflejen la información de los datos originales y con mayor independencia entre ellos.

Geométricos o basados en grafos: Estos son tipos de características que usan algún tipo de representación gráfica generalmente la teoría de grafos ya que con estos se pueden representar muchos

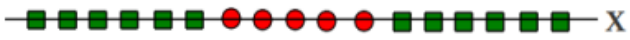
problemas importantes como redes sociales, interacciones entre bacterias, redes de computadoras, etc. Sobre estas representaciones se pueden aplicar diferentes técnicas de la teoría de grafos para determinar cuales son los mejores descriptores como por ejemplo la técnica de detección de comunidades en la cual se busca agrupar nodos fuertemente conectados en clusters mientras que los nodos que están en diferentes clusters son los que están mas débilmente conectados. Estos tipos de técnicas varían según si se aplican a grafos dirigidos, no dirigidos, con pesos, sin pesos, etc. usando métricas como la distancia, centralidad, densidad, además del ya mencionado nivel de conexión entre nodos.

Espectrales o basadas en series de tiempo: Estos son tipos de características que se derivan de datos en secuenciales o continuos como por ejemplo una serie de eventos que ocurren uno detrás del otro. Se puede tomar en cuenta la periodicidad, magnitud del espectro, variaciones temporales, etc. y se usan técnicas como la transformada de Fourier.

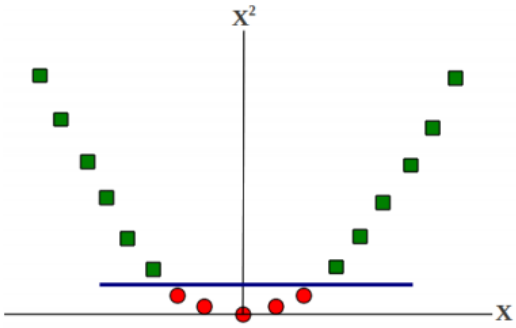
Otra forma de clasificar las técnicas usadas en feature extraction es en métodos lineales o no lineales:

Métodos lineales: usualmente los problemas de clasificación pueden resolverse mediante separaciones lineales en un plano o hiperplano, es decir, las clases pueden delimitarse fácilmente "trazando" estas lineas que las separen. Entre estas técnicas se encuentran la ya mencionada PCA y también el Análisis de Discriminantes Lineales (LDA) el cual es un método que también busca reducir dimensionalidad pero a diferencia de PCA que es no supervisado (no toma en cuenta clases) LDA es supervisado y funciona buscando proyecciones óptimas que maximicen las distancias entre las clases y minimicen distancias de los datos dentro de esas clases.

Métodos no lineales: cuando la clasificación es mas compleja y las clases no son linealmente separables, es decir, los límites entre ellas son discontinuos, se utilizan métodos no lineales ya que ayudan a realizar esta separación de clases. Entre estos métodos pueden usarse técnicas como realizar un mapeo de los datos a dimensiones mas altas donde sí sea posible realizar una separación lineal y luego aplicar un modelo lineal. Como ejemplo se puede considerar la siguiente figura



en ella se aprecia como las dos clases representadas (cuadros verdes y círculos rojos) no son linealmente separables. Se puede realizar un mapeo sencillo a dos dimensiones mediante la transformación $X \rightarrow \{X, X^2\}$ resultando en un doble valor por cada uno del ejemplo original que al representarlo en un plano se puede apreciar que sí es linealmente separable en esa nueva representación



Entre estos tipos de métodos están las redes neuronales multicapa, k-NN, SVM no lineales, etc.