**\section{Entrenamiento utilizando PCA}**

Para el uso de PCA, a pesar de que ya se tiene una noción de que clasificadores funcionan mejor en este tipo de problemas, igualmente se analiza para los otros tipos de clasificadores, ya que de esta manera se puede evaluar el efecto de aplicarlo en estos, además sirven como guía para el análisis posterior.

Lo primero a determinar claramente, es el número de componentes principales que deben ser utilizadas para disminuir los tiempos de procesamiento y el número de componentes no ortogonales, es decir, reducir la información redundante total. Para ello se debe seleccionar el mínimo número de componentes principales, ya que mientras más se añaden, es mucho más probable que los datos presenten información no relevante, es decir, ruido e información duplicada. Obviamente, es esencial reducir este tipo de datos para así mejorar el desempeño de los algoritmos de clasificación.

Para la selección, es necesario realizar pruebas que ayuden a la toma de decisiones sobre el número de componentes principales a utilizar, ya que hasta el momento no existe un algoritmo que lo determine automáticamente, por lo que depende netamente del experimentador y los resultados deseados en términos de accuracy y ahorro de procesamiento. Para ello se proponen tres métodos explicados a continuación. El primer método se basa en la información contextual presente en cada componente principal a través de los valores propios de la matriz de covarianzas. Estos valores propios representan que tanto se explica la varianza en una determinada dirección, en donde la dirección es la determinada por el vector propio asociado al valor propio de esa componente principal. Por lo anterior, a medida que los valores propios de una determinada componente principal incrementan su valor, esta componente principal tiende a tener más información valiosa en términos generales para el conjunto total de datos.

Después de aplicar el algoritmo PCA, el espacio de características se mantiene constante según el número de componentes originales, por lo que si el radiomap se proyecta directamente no hay mejoría y la complejidad computacional incluso puede aumentar. Para la selección, se incorpora una variable $U$, donde esta representa el número de componentes principales seleccionadas. Como se explica anteriormente, el primer método es determinar los valores propios que aportan más información(varianza), mientras más alto su valor, la componente principal asociada mantiene más información valiosa para los datos, es decir, puede interpretarse como la importancia de esa componente. Para determinar los valores propios más relevantes, estos se determinan según un criterio muy conocido, el cual es seleccionar aquellos valores propios con valor cercano o mayor a 1. Para ello, se grafican los valores propios en orden descendente:

\begin{figure}[ht!]

\centering

**\includegraphics**[width=.6\textwidth]{figures/eigenvalues.png}

\caption[Análisis de los valores propios obtenidos a partir de la matriz de covarianza]{Análisis de los valores propios obtenidos a partir de la matriz de covarianza utilizada en el algoritmo PCA \\

{\scriptsize (Fuente: Elaboración Propia)}}

**\label{fig:manifold}**

\end{figure}

Como se observa, los primeros 4 valores propios cumplen la heurística de selección necesaria, por lo que mantienen la mayor parte de la información del \textit{dataset}.

Para el segundo método, es necesario establecer la suma acumulada porcentual de la varianza explicada, lo cual es sinónimo de el porcentaje de información relevante retenida o acumulada. Para ello, es necesario igualmente utilizar los valores propios que representan la varianza para la selección. La fórmula a utilizar entonces es:

\begin{equation} \frac{\sum\_{i=1}^{U}\lambda\_{i}}{\sum\_{i=1}^{M}\lambda\_{i}} > \xi

\end{equation}

De donde $\{ \lambda\_{1}, \lambda\_{2}, ..., \lambda\_{M}\}$ corresponden a los valores propios y :

\begin{equation}

\frac{\sum\_{i=1}^{U}\lambda\_{i}}{\sum\_{i=1}^{M}\lambda\_{i}}

\end{equation}

Representa la suma acumulada que retiene la información. Además, el limite requerido es $\xi$. Este límite representa la cantidad de información mínima con las cual se acepta el criterio, y habitualmente se elige un valor entre 70 a 80\%. Para esto se muestra la \autoref{fig:varianza\_ratio} la cual describe esta situación.

\begin{figure}[ht!]

\centering

**\includegraphics**[width=.6\textwidth]{figures/varianza\_ratio.png}

\caption[Proporción de la varianza explicada PCA]{Proporción de la varianza explicada respecto a la varianza total para cada componente ordenada por su valor en orden descendente \\

{\scriptsize (Fuente: Elaboración Propia)}}

**\label{fig:varianza\_ratio}**

\end{figure}

Como se aprecia en la imagen anterior, las primeras 5 componentes principales retienen la cantidad de información especificada según el límite impuesto $\xi =0.8$. Para clarificar este punto se construye una tabla resumen con los valores propios, la proporción de varianza explicada y la suma acumulada de estas proporciones.

\begin{table}[]

\centering

\caption[Tabla resumen varianza explicada y acumulada PCA]{Tabla resumen varianza explicada y acumulada PCA}

**\label{my-label}**

\begin{tabular}{|c|c|c|c|c|c|c|c|c|}

\hline

Indicador & PC1 & PC2 & PC3 & PC4 & PC5 & PC6 & PC7 & PC8 \\ \hline

Valor Propio & 2.1271 & 1.3665 & 1.3154 & 0.92088 & 0.7652 & 0.7058 & 0.4637 & 0.3363 \\ \hline

Proporción Varianza Explicada & 0.2658 & 0.1707 & 0.1644 & 0.1150 & 0.0956 & 0.0882 & 0.0579 & 0.0420 \\ \hline

Suma Acumulada Proporción & 0.2658 & 0.4366 & 0.6010 & 0.7161 & 0.8117 & 0.9000 & 0.9579 & 1.000 \\ \hline

\end{tabular}

\end{table}

Con respecto al tercer método, este se basa en seleccionar las primeras componentes principales según los resultados obtenidos en los clasificadores, por lo que se necesita PCA ya aplicado, no solo basta con la matriz de correlación, si no que los algoritmos completos. Para realizar este método entonces se decide aplicar PCA en todos los algoritmos de clasificación nombrados anteriormente, con ellos se pretende establecer los mejores valores de accuracy y en qué puntos a nivel de componentes principales, la mayoría de los clasificadores establecen sus mejores puntuaciones en términos de accuracy, por lo que añadir más componentes no provoca una mejoría significativa.

La \autoref{fig:comparativa\_pca} muestra la comparativa de los clasificadores.

\begin{figure}[ht!]

\centering

**\includegraphics**[width=.6\textwidth]{figures/comparativa\_clasificadores\_pca.png}

\caption[Comparativa de resultados obtenidos por los clasificadores utilizando PCA]{Comparativa de resultados obtenidos por los clasificadores utilizando PCA \\

{\scriptsize (Fuente: Elaboración Propia)}}

**\label{fig:comparativa\_pca}**

\end{figure}

Según la comparativa, es claro que la mayoría de los clasificadores presentan una gran mejoría hasta la quinta componente, y posteriormente la ganancia no es significante respecto a la información aportada a los respectivos algoritmos. Esto indica al igual que los otros métodos de selección, que a partir de la quinta componente la varianza aportada no es realmente significativa, por lo que puede reducirse la dimensionalidad de los datos a 5, ya que los tres métodos utilizados presentan resultados similares y coinciden en este valor de componentes principales.

Posteriormente, se procede a evaluar según el procedimiento realizado anteriormente para los clasificadores sin utilizar PCA. Los resultados obtenidos en cada clasificador se muestran en la \autoref{fig:comparativa\_clasificadores\_pca}. Se debe destacar que los hiperparámetros de cada algoritmo no fueron cambiados respecto a sus versiones sin utilizar PCA.

\begin{figure}[ht!]

\centering

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/NB\_PCA.png}

\caption{Naive Bayes PCA}

**\label{fig:sub1}**

\end{subfigure}%

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/SVM-RBF-PCA.png}

\caption{SVM RBF PCA}

**\label{fig:sub2}**

\end{subfigure}

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/SVM-Lineal-PCA.png}

\caption{SVM Lineal PCA}

**\label{fig:sub1}**

\end{subfigure}%

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/knn-results-PCA.png}

\caption{K-NN PCA}

**\label{fig:sub2}**

\end{subfigure}

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/decision-tree-PCA.png}

\caption{Decision Tree PCA}

**\label{fig:sub1}**

\end{subfigure}%

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/random-forest-PCA.png}

\caption{Random Forest PCA}

**\label{fig:sub2}**

\end{subfigure}

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/adaboost-PCA.png}

\caption{Adaboost PCA}

**\label{fig:sub1}**

\end{subfigure}%

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/qda-PCA.png}

\caption{QDA PCA}

**\label{fig:sub2}**

\end{subfigure}

\caption[Curvas de aprendizajes obtenidas por cada clasificador utilizando PCA]{Curvas de aprendizajes obtenidas por cada clasificador utilizando PCA con 5 componentes\\

{\scriptsize (Fuente: Elaboración Propia)}}

**\label{fig:comparativa\_clasificadores\_pca}**

\end{figure}

En este caso, se obtienen resultados similares a los obtenidos sin utilizar PCA, esto es un indicativo de que la elección de las componentes principales es correcta. Para analizar las redes neuronales utilizando PCA se procede a graficar la accuracy y la función de perdida al igual que en casos anteriores.

\begin{figure}[ht!]

\centering

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/nn\_plot\_pca.png}

\caption{Accuracy PCA}

**\label{fig:sub1}**

\end{subfigure}%

\begin{subfigure}{.5\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/nn\_loss\_pca.png}

\caption{Loss PCA}

**\label{fig:sub2}**

\end{subfigure}

\caption[Accuracy y Loss(Costo) obtenidos en el entrenamiento de la red neuronal profunda con dos capas ocultas utilizando PCA]{Accuracy y Loss(Costo) obtenidos en el entrenamiento de la red neuronal profunda con dos capas ocultas utilizando PCA con 5 componentes principales. \\

{\scriptsize (Fuente: Elaboración Propia)}}

**\label{fig:nn\_metrics\_pca}**

\end{figure}

Al igual que los clasificadores convencionales, los resultados son muy similares a los obtenidos mediante clasificadores sin utilizar PCA. Esto es un indicativo de que la elección de las componentes principales es acertada. Para corroborar esto se construye la tabla \autoref{tabla-pca} la cual es símil a la \autoref{tabla-clasificacion}

\begin{table}[ht!]

\centering

\caption{My caption}

**\label{tabla-pca}**

\begin{tabular}{|c|c|c|c|c|}

\hline

Algoritmo & Accuracy & Error medio X & Error medio Y & Error Absoluto \\ \hline

NN & 93\% & 1.8956 & 0.6589 & 2.0068 \\ \hline

$SVM(RBF, C=1, \gamma = 4)$ & 92.39\% & 2.4581 & 0.7830 & 2.5797 \\ \hline

$KNN(k = 2)$ & 90.83\% & 2.1381 & 0.4872 & 2.1929 \\ \hline

QDA & 71.25\% & 15.9418 & 9.3042 & 18.4583 \\ \hline

$SVM(Lineal, C=1)$ & 66.20\% & 19.5878 & 10.4824 & 22.2162 \\ \hline

Random Forest & 64.15\% & 21.3284 & 5.4472 & 22.0130 \\ \hline

Decision Tree( max depth = 5) & 56.96\% & 33.5151 & 8.9163 & 34.6808 \\ \hline

Naive Bayes & 50.71\% & 31.3406 & 10.0727 & 32.9194 \\ \hline

Adaboost & 32.20\% & 73.9345 & 9.3042 & 74.5176 \\ \hline

\end{tabular}

\end{table}

Los valores obtenidos entonces demuestran que a pesar de que la accuracy se reduce solo un poco, los valores de error medio decrementan significativamente en el contexto del problema abordado, ya que para el posicionamiento en interiores se espera un error lo más pequeño posible, ya que esto significa una mejor localización. Por lo anterior, se debe tener en cuenta estos factores al momento de seleccionar las componentes principales, ya que, al perder información, se pierde exactitud en las mediciones, lo que repercute significativamente al momento de realizar pruebas reales.

El comportamiento en términos generales sigue el mismo patrón que al no aplicar técnicas de reducción de dimensionalidad, es decir, los clasificadores mantienen el orden relativo en cuanto a accuracy y nuevamente los clasificadores capaces de distinguir patrones no lineales en los datos son los dominantes, lo cual prueba estas características de los datos. También se debe notar que los tres primeros lugares corresponden nuevamente a redes neuronales profundas, máquina de vectores de soporte con kernel RBF y finalmente vecinos más cercanos con $k=2$.

Por lo anterior, se decide utilizar los tres clasificadores NN, SVM y KNN para realizar la experimentación en el estacionamiento descrito anteriormente. A continuación, se define el método utilizado para realizar las pruebas.

**\section{Fase Online}**

Para realizar las pruebas dentro del recinto es necesario establecer el marco a utilizar para determinar la exactitud y precisión del sistema. En primer lugar, se debe tener en cuenta que no hay forma de determinar el error absoluto, producto de que para ello se debe proporcionar la posición real, lo cual es contradictorio, ya que esto es precisamente lo que se busca. Existen métodos de buscar este error, por ejemplo \citep{Ugave} describe una forma de medir los errores mediante el uso de la creación de caminos predefinidos. Luego para medir la precisión, se infiere la posición estimada cada 10 metros siguiendo la trayectoria definida. Luego, con la posición estimada y la posición teórica determinada por el camino establecido, se puede obtener la distancia que representa el error, la cual está definida mediante el vector perpendicular desde la trayectoria definida hasta el punto predicho.

Otro tipo de enfoque es utilizar el tiempo recorrido y mediante las ecuaciones de movimiento obtener una posición aproximada que puede ser comparada con la inferida por los algoritmos de máquinas de aprendizaje.

El problema con estos enfoques es que se basan en propiedades del experimentador, como por ejemplo utilizar sensores basados en detección de pasos, lo cual no es del todo preciso, o también al utilizar las ecuaciones de posición se requiere que la velocidad del experimentador sea constante y medir exactamente en un instante indicado, lo cual dificulta mucho las pruebas y también agrega factores de error externos lo que se traduce en resultados ruidosos y poco fidedignos. Entonces lo que se propone para determinar los resultados son dos formas llamadas método estático y método dinámico. En el método estático lo que se hace es permanecer quieto en un determinado punto durante un tiempo predefinido. El tiempo utilizado en este caso corresponde a 15 minutos, y el punto es especifico es $(22, 12)$. Para el caso del método dinámico, lo que se busca es abarcar la mayor cantidad de puntos posibles y establecer de ante mano el punto en el cual se realizaran las mediciones a través de una caminata. En este caso se decide hacer una caminata a través de todos los 44 puntos posibles de izquierda a derecha, arriba a abajo durante un tiempo de 30 minutos.

Para realizar las pruebas se adiciona una pantalla a la aplicación precisamente para la fase Online, la \autoref{fig:fase\_online} muestra esta situación.

\begin{figure}[ht!]

\centering

\begin{subfigure}{.3\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/fase\_online1.png}

\caption{Selección algoritmo}

**\label{fig:online1}**

\end{subfigure}%

\begin{subfigure}{.3\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/fase\_online2.png}

\caption{Elección patrón}

**\label{fig:online2}**

\end{subfigure}

\begin{subfigure}{.3\textwidth}

\centering

**\includegraphics**[width=.8\linewidth]{figures/fase\_online3.png}

\caption{Posicionamiento real}

**\label{fig:online3}**

\end{subfigure}

\caption[abs]{Pantalla utilizada en la fase Online para la realización de las pruebas de experimentos reales \\

{\scriptsize (Fuente: Elaboración Propia)}}

**\label{fig:fase\_online}**

\end{figure}

Como se observa en la imagen \autoref{fig:fase\_online}, en primer lugar se debe elegir el algoritmo a utilizar en el posicionamiento, luego seleccionar el patrón para elegir el punto en donde se realizaran las pruebas y finalmente iniciar, lo cual comienza a posicionar al usuario en tiempo real. Para el método dinámico los pasos son igual, solo se debe utilizar el \textit{checkbox} al lado del botón iniciar y cambiar progresivamente el punto a utilizar mientras se sigue el camino de predicción.

A pesar de que se selecciona un algoritmo para la visualización del posicionamiento, internamente la aplicación ejecuta todos los algoritmos mostrados, es decir NN, SVM y KNN con sus respectivas versiones utilizando PCA. El fin de esto es ahorrar tiempo, ya que de esta forma se pueden obtener los resultados para todos los algoritmos concurrentemente sin necesidad de ejecutarlos de manera secuencial. Luego la aplicación tiene la capacidad de crear archivos de \textit{log}, es decir, guardar los resultados obtenidos. El formato de estos archivos es diseñado para su posterior análisis. La primera línea contiene el tipo de método, estático o dinámico. La siguiente línea de texto muestra el punto teórico con el cual se comparan los datos. Posteriormente el nombre del algoritmo, luego una lista de valores de la posición $X$, a continuación, una lista de valores $Y$ y finalmente el tiempo de ejecución del algoritmo. Para cada algoritmo se repite esta estructura, en el orden mostrado en la \autoref{fig:online1}. Claramente la lista de valores $X$ e $Y$ tienen las mismas dimensiones y forman las predicciones finales según su posición relativa, es decir, el primer elemento de la lista $X\_{1}$ con el primer elemento de la lista $Y\_{1}$ forman el primer punto predicho por el algoritmo $(X\_{1}, Y\_{1})$.

Para clarificar esto, la \autoref{online\_results} muestra un extracto de un archivo obtenido luego de realizar la experimentación de la fase online. Este archivo es de tipo dinámico, con lo que contiene múltiples puntos de clasificación.

\begin{figure}[ht!]

\centering

**\includegraphics**[width=.6\textwidth]{figures/online-results.png}

\caption[Archivo obtenido en la experimentación online]{Archivo obtenido en la experimentación online \\

{\scriptsize (Fuente: Elaboración Propia)}}

**\label{fig:online\_results}**

\end{figure}