**Introducción**

Análisis de regresión con técnicas de aprendizaje supervisado

La Minería de Datos o Data Mining (DM) agrupa a un conjunto de herramientas y métodos orientados al descubrimiento de conocimiento a partir de bases de datos (KDD), del inglés Knowledge Discovery from Databases. KDD integra elementos como métodos estadísticos, aprendizaje de máquina y visualización de datos.

Dentro del área de aprendizaje de máquina el aprendizaje supervisado es una técnica que consiste en aprender una función a partir de un conjunto de datos de entrenamiento. El conjunto de entrenamiento consiste en pares de objetos de entrada y salida esperada. Utilizando estos datos se aprenderá una función que luego será capaz de predecir la salida esperada utilizando datos de entrada nunca vistos.

Dentro del aprendizaje supervisado existen diferentes métodos de predicción, los cuales son más o menos eficientes de acuerdo a la naturaleza de los datos del problema en cuestión. Ejemplos de ellos son las maquinas de soporte vectorial, redes neuronales, árboles de decisión, etc. [1]. Estos algoritmos son utilizados en diversas áreas, incluyendo agricultura, finanzas, historia, industria, relaciones laborales y ciencias del medioambiente, entre otras.

Cuando la salida de la función aprendida es un valor continuo la tarea de aprendizaje se denomina regresión. El análisis de regresión es un conjunto de técnicas de modelado y análisis numérico la cual consiste en determinar el valor de una variable dependiente a partir de un conjunto de variables independientes. Esta técnica es utilizada para realizar pronósticos, inferencia, prueba de hipótesis y evaluación de causalidad entre otras cosas.

Usualmente se ha utilizado a este efecto la regresión lineal, pero los modelos obtenidos sólo operan con atributos numéricos e imponen una dependencia puramente lineal. Otros métodos de regresión son los provenientes del área de aprendizaje supervisado como los árboles de regresión y las redes neurales para regresión. Los árboles de regresión [3] se generan de forma similar a los de decisión, con valores promedio en cada hoja. El modelo neuronal de regresión generalizada (GRNN), que fue propuesto y desarrollado inicialmente por Specht [2], posee la propiedad deseable de no requerir ningún entrenamiento iterativo, es decir, puede aproximar cualquier función arbitraria entre vectores de entrada y salida, realizando la estimación de la función directamente a partir de los datos de entrenamiento.

Nuestra propuesta consiste en la evaluación de un conjunto de algoritmos de regresión provenientes del área de aprendizaje supervisado y su comparación con modelos de regresión clásicos en el marco de un caso de estudio específico. A partir de los resultados experimentales que se obtengan de estos algoritmos, se seleccionará el método de mejor performance para ser usado en un entorno real para el problema de predicción que se abordará.

**Caso de estudio**

En la actualidad existen una variedad de deportes acuáticos que se ven beneficiados por los pronósticos acerca del oleaje en una zona cercana a la costa. Ejemplos de estos pueden ser el surf, windsurf, kitesurf, kayak, etc. Los deportistas están sumamente atentos a las condiciones que presenta el mar diariamente para evaluar la playa en la que van a practicar su deporte. Utilizando un pronóstico de oleaje el deportista puede planificar sus actividades para lograr estar presente en el lugar que presenta las mejores condiciones. Eventualmente el deportista puede evitarse recorridas extensas buscando el sitio con las mejores condiciones del día.

Actualmente los pronósticos de oleaje son derivados a partir de la resolución de un modelo matemático global sumamente complejo cuyas predicciones son calculadas en zonas alejadas de la costa, varios kilómetros mar adentro. Este provee de la predicción de la altura de las olas, la dirección de las mismas y el período (tiempo entre una ola y la siguiente). Si bien este pronóstico tiene buen rendimiento en alta mar, no sucede lo mismo en la cercanía de la costa. Esto se debe a que la interacción de las olas con el lecho marino cercano no es contemplada por estos modelos. De acuerdo a la geografía de la costa puede suceder que en un rango de 15 km un lugar presente condiciones apropiadas para la práctica del deporte mientras que en otro las condiciones sean totalmente desfavorables. Por esta razón es que la utilización de un pronóstico preciso en las cercanías de la costa (exactamente donde el mar es lo suficientemente poco profundo para que las olas rompan) es de suma utilidad.

Existen modelos para contemplar el comportamiento de las olas en las cercanías de la costa. Los mismos reciben como entrada los datos de pronóstico de alta mar e información acerca del lecho marino. Utilizando estos datos los mismos proveen un refinamiento al pronóstico inicial. Si bien estos modelos poseen un buen comportamiento, su limitación esta relacionada con la obtención de los datos del lecho marino. Usualmente la obtención de estos datos es costosa y no se encuentra disponible en todos los lugares.

Existen algunos trabajos previos que enfocan el tema de predicción de oleaje [4,5]. En ambos se utilizan datos de alta mar junto con observaciones visuales, relevadas por un observador experto, de la altura de las olas en las proximidades de la costa. Con estos datos se entrena una red neuronal que pueda estimar la altura de la ola en la cercanía de la costa. En [4] los datos pertenecen a la costa australiana. Se entrena una red neuronal del tipo feed-forward y se realiza una comparación entre un método numérico para refinamiento del modelo de alta mar y un método de redes neuronales, dando este ultimo una mejor precisión en los pronósticos. En [5] los datos son de la costa norte de Irlanda. Se experimenta con un ensamble de redes neuronales del tipo MDN (mixture density networks) y la salida de la red neuronal representa una distribución de probabilidad de la altura de la ola en la cercanía de la costa. Los resultados encontrados fueron positivos haciendo el modelo de red neuronal desarrollado útil para la estimación del tamaño de la ola en la cercanía de la costa.

**Trabajo realizado**

Nuestro estudio consistió en utilizar modelos de aprendizaje de máquina para aprender de casos pasados como un determinado pronóstico de alta mar se vio reflejado en las cercanías de la costa. De esta manera ante un nuevo pronóstico oceánico se logró predecir cuál fue el efecto del mismo en las cercanías de la costa. Nos centramos en la predicción de la altura de la ola en la cercanía de la costa, mejorando la predicción del pronóstico de alta mar. Para realizar esta predicción utilizamos métodos de regresión los cuales utilizaron como datos de entrada: el pronóstico de alta mar y la observación de la altura en la cercanía de la costa.

Los métodos de regresión utilizados fueron:

* Regresión lineal
* Redes neuronales
* Maquinas de soporte vectorial
* Arboles modelos y de regresión

Evaluamos el desempeño de cada modelo sometiéndolos a un entrenamiento con datos reales (tanto de lecturas de altamar como de observaciones visuales costeras) y luego evaluando las predicciones con un conjunto datos diferentes a los utilizados para entrenar cada clasificador (datos nuevos que nunca antes han usado para entrenarse). Finalmente implementamos un sistema de predicción de olas en la cercanía de la costa en cualquier lugar del planeta utilizando como algoritmo de regresión aquel que presentó los mejores resultados en cada una de las diferentes pruebas a las que se sometieron en la etapa de experimentación (Ver capitulo 4 - Evaluaciones).

La evaluación experimental se realizó utilizando datos históricos de las condiciones de alta mar y de observaciones visuales correspondientes a la isla de Oahu, Hawái. Los datos de alta mar se obtuvieron de un archivo histórico de las salidas del modelo numérico WaveWatch III, provistos por la agencia americana NOAA (National Ocenographic and Atmospheric Administration), comprendiendo la altura, dirección y periodo de ola, dirección y velocidades de los vientos. Los datos de las observaciones visuales usados solo brindan la altura de la ola vista desde la costa. Las observaciones han sido relevadas en su mayoría por el servicio de guardavidas de la isla de Oahu y se encuentran disponibles para cinco playas diferentes: Sunset, Diamond Head, Makapu, Makaha y Ala Moana ubicadas alrededor de la isla. Tanto de la información de alta mar como de las observaciones visuales se cuenta con más de 6 años de registros con una periodicidad diaria.

Dado que las lecturas de WAVEWATCH III se actualizan cada 3 horas, tanto durante el día como la noche, y debido a que las observaciones visuales se han realizado en horarios en que se cuenta con luz solar, el primer filtro aplicado a la información que ofrece el modelo WAVEWATCH III fue eliminar todas las lecturas que no correspondían con un horario de luz natural en Oahu. Por otra parte, las observaciones visuales nos indicaban la altura de la ola más grande que arribo a la costa en el día, por esto el segundo filtro aplicado a las lecturas de WAVEWATCH III fue dejar únicamente la lectura diaria con la mayor altura de ola. De esta forma acoplamos el estado de alto mar brindado por el modelo WAVEWATCH III y la observación de la altura de la ola registrada en la costa.

Para llevar a cabo el entrenamiento de los clasificadores utilizamos la herramienta Weka. Esta herramienta provee los algoritmos de regresión (entre otros), utilizados en nuestro estudio durante la etapa de experimentación. Weka recibe como entrada un conjunto de instancias descriptas a través de una serie de atributos. Cada instancia se compone un conjunto de variables predictoras y uno o más variables a predecir. En nuestro estudio las variables predictoras son las lecturas del modelo WaveWatch III (altura de ola, periodo, velocidad del viento, dirección de ola) y el atributo de clase o variable a predecir fue la observación visual en la costa. Entonces dado un conjunto de instancias como entrada, y un algoritmo de regresión, Weka logra entrenar un clasificador que es capaz de recibir únicamente instancias con las variables predictoras y devolver por cada una el valor del atributo de clase a predecir.

La clave del estudio fue encontrar la forma de armar las instancias en base a la información de la cual disponíamos y al mismo tiempo buscar y configurar el algoritmo de regresión que mejores resultados mostraba al entrenarse con dichas instancias (es decir el que lograba predecir las olas en la ubicación del observador minimizando el margen de error y maximizando la correlación entre los datos de las lecturas de alta mar y la predicha por el clasificador).

En este trabajo se experimentó con numerosos modelos de armado de instancias. Para cada modelo de instancia se armaron diferentes conjuntos de entrenamiento, los cuales fueron utilizados como entrada en los diferentes algoritmos de regresión. La combinación de variables predictoras (instancias) que mejores resultados mostraron en los experimentos fue la siguiente:

Modelo de instancia escogido:

* altura significante de ola actual (variable predictora WW3)
* dirección promedio actual (variable predictora WW3)
* período promedio actual (variable predictora WW3)
* altura significante de ola 1 día antes (variable predictora WW3)
* dirección promedio 1 día antes (variable predictora WW3)
* período promedio 1 día antes (variable predictora WW3)
* altura significante de ola 2 días antes (variable predictora WW3)
* dirección promedio 2 días antes (variable predictora WW3)
* periodo promedio 2 días antes (variable predictora WW3)
* observación visual (atributo de clase, histórico de observaciones de Oahu)

Este modelo utilizado se basó en armar cada instancia para un determinado GridPoint en una determinada fecha con las lecturas del WAVEWATCH III correspondientes a la ola más grande registrada ese día en un horario con luz solar, junto con la observación visual de altura de ola de ese mismo día, pero además con las lecturas de WAVEWATCH III correspondientes a la ola más grande registrada un día antes y dos días antes (ambas durante un horario con luz solar). De esta manera armamos instancias con nueve variables predictoras de WAVEWATCH III y un atributo de clase que representa la observación visual.

Por otra parte el algoritmo de regresión elegido en base al rendimiento del mismo fue Maquinas de Soporte Vectorial, cuyos resultados al ejecutarlos con las instancias descriptas anteriormente pueden observarse en el Capitulo 4 – Evaluaciones.

Con base en este clasificador entrenado desarrollamos un sistema web denominado “Surf-Forecaster” el cual es capaz, si se le suministra la información necesaria (histórico de observaciones visuales de cualquier lugar deseado), de dar un pronóstico con hasta una semana de anticipación en incrementos de tres horas, especificando la altura de las olas, dirección de ola, periodo de ola y dirección y velocidad del viento exactamente en la ubicación que al usuario le interesa, mejorando ampliamente los pronósticos existentes los cuales brindan esta información pero en zonas muy alejadas a la costa, que por lo general reportan el mismo pronostico para toda una región o localidad, la cual pierden precisión en sitios con una geografía muy irregular o donde el tipo de fondo y profundidad del mar (entre otros aspectos) cambian vertiginosamente entre lugares separados por tan solo unos pocos kilómetros.

El sistema desarrollado puede ser de gran utilidad tanto para deportistas que diariamente recorren una localidad en busca de las mejores condiciones para su deporte como asi también por ejemplo para tomar medidas de prevención anticipadas en lugares donde olas de gran tamaño pueden llegar a alcanzar a la población de ese lugar.

Surf-Forecaster ofrece además la posibilidad de registrase en el sistema (Ver sección 5.2 – Tipos de usuario). Al ser un usuario registrado el sistema le permitirá dar de alta sus olas preferidas (en cualquier lugar del mundo), generar comparaciones entre ellas y persistirlas, para no tener que volver a crearlas cada vez que quiere comparar las mismas olas.

Referencias

[1] Witten, I. H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann, 2005.

[2] Specht, D. F.: A General Regression Neural Network. IEEE Transactions on Neural Networks. Vol. 2(6), pp. 568-576. 1991.

[3] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone.. Classification and regression trees. Monterey, California, U.S.A.: Wadsworth, Inc. 1984.

[4] Browne, M. Strauss, D. Castelle, B. Blumenstein, M. Tomlinson, R. Lane, C.: Empirical Estimation of Nearshore Waves From a Global Deep-Water Wave Model. IEEE Geoscience and Remote Sensing Letters. Vol. 3(4), pp. 462-466. 2006.

[5] Carney, M., Cunningham, P., Dowling, J. and Lee, C.: Predicting probability distributions for surf height using an ensemble of mixture density networks. In Proceedings of the 22nd International Conference on Machine Learning. Vol. 119, pp. 113-120. 2005.