Diseño de un modelo predictivo para equipos profesionales en Dota 2

Anteproyecto

Línea de investigación del grupo FICB-PG: Línea de Investigación en educación y tecnología

Rene Felipe Cardozo 0321040262

Asesor: Javier Niño

Diciembre 2018

_____

Resumen

La investigación de esta tesis consiste en explorar los datos de las partidas del juego Dota 2, donde se analizarán las partidas individuales de los jugadores pertenecientes a un equipo profesional y en las partidas realizadas en ligas mayores y premier. Por un lado, analizará las partidas individuales de un jugador profesional, entrenando un algoritmo de aprendizaje, para obtener un modelo predictivo. Por otro lado, se implementará un árbol de decisión que será a su vez alimentado por el anterior modelo aplicando este algoritmo en el campo de las ligas mayores y premier, para predecir un equipo ganador.

Palabras clave: Machine learning, modelo predictivo, minería de datos, aprendizaje supervisado.

Abstract

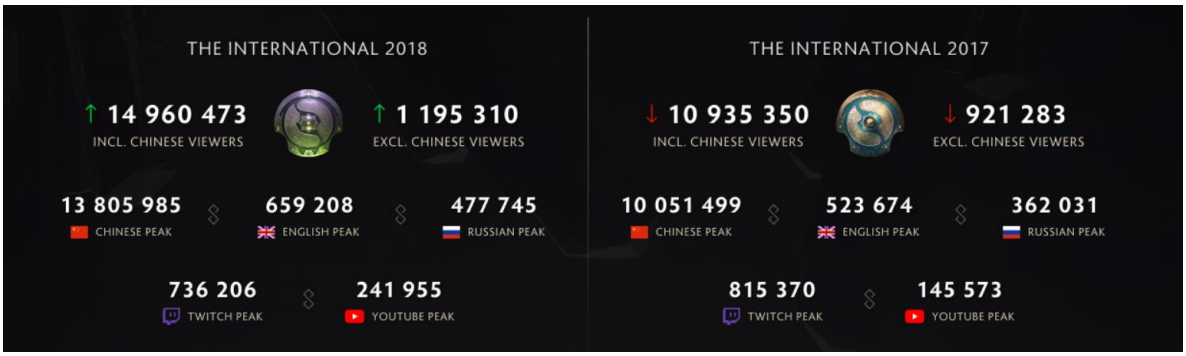
The research that will be carried out in this thesis will consist in exploring the data of the games of the game Dota 2, where the unique games of the players belonging to a professional league will be found, and in the games made in the major and premier leagues. On the one hand, we will analyze the unique games of a professional player, training a series of supervised learning algorithms and obtain a predictive model. On the other hand, a decision tree will be implemented that will be fed by the previous model, applying this tree in the field of the major and premier leagues, to predict a winning team.

Key words: Machine learning, predictive model, data mining, supervised learning.

INTRODUCCIÓN

Según Wagner la definición de esport es “un área de actividades deportivas en la que las personas desarrollan y entrenan habilidades mentales o físicas con el uso de las tecnologías de la información y la comunicación” [1].

Abro la introducción con la definición dada por Wagner sobre esports, porque los esports son considerados muchas veces solo como un entretenimiento. Sin embargo, existen datos estadísticos como los de ‘The Motley Fool’ que demuestran que los esports ya han alcanzado grandes audiencias, como es el caso de League of leguends con 14.7 millones de espectadores recurrentes, teniendo en cuenta que el juego 7 de la NBA en el 2016 fue de 44.5 millones de espectadores [2]. Una cifra considerable teniendo en cuenta que la de league of leguends o su juego rival Dota 2 sigue creciendo.



[3]

¿Qué es Dota 2?

Dota 2 es un juego gratuito **Multiplayer Online Battle Arena (MOBA)** desarrollado por la corporación Valve. Consta de dos equipos llamados Dire y Radiant, cada equipo se conforma por 5 héroes diferentes, gana el equipo que destruya el edificio “ancient” del otro equipo. No existe un límite de tiempo para lograr destruir el edificio objetivo “ancient” y su mapa al igual que muchos otros MOBA como *League of Legends* o *Heroes of the Storm*, se basa en tres líneas protegidas con torres y para destruir el último edificio es necesario primero destruir estas torres en las diferentes líneas, existen cerca de 215 héroes elegibles, cada uno con diferentes poderes y habilidades.

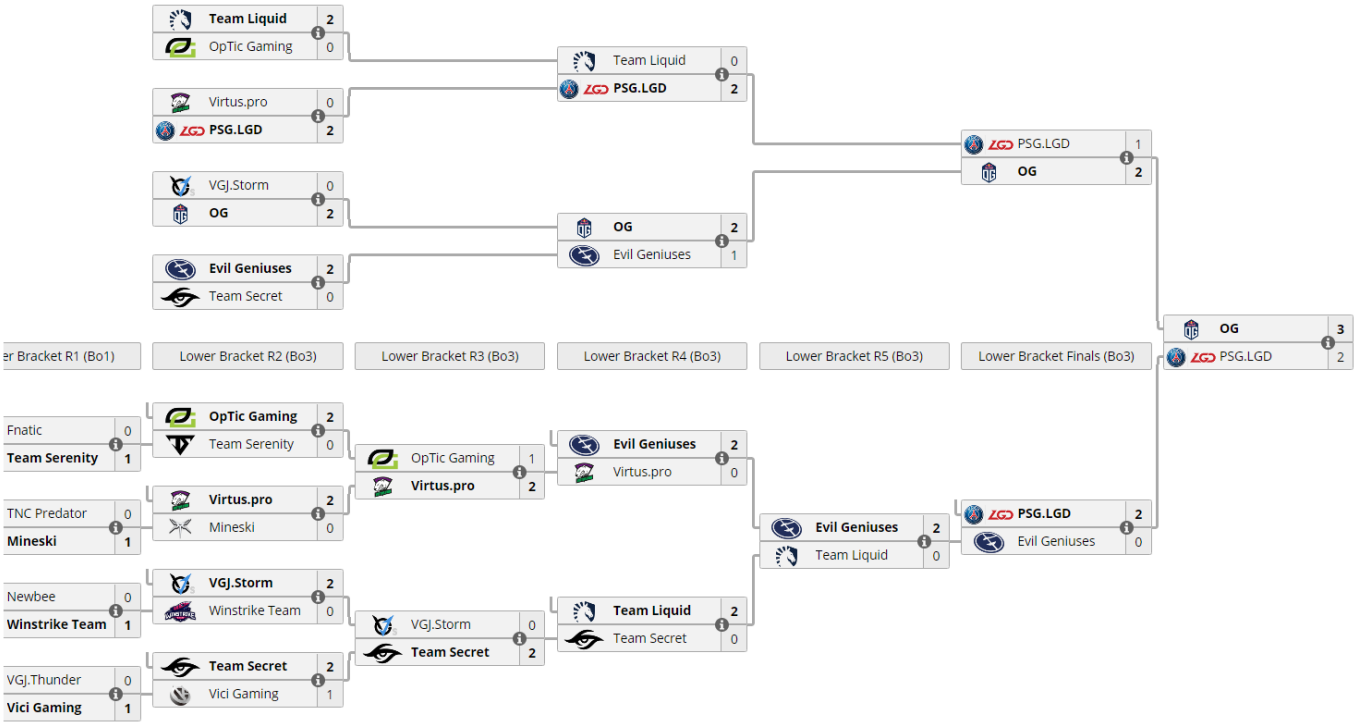
[4]

Equipos profesionales

En Dota 2 existen ligas de diferentes tipos, existen las más grandes que son la premier y las ligas mayores, en las cuales los equipos élite compiten para llevarse un gran premio, su liga de mayor representación es “The International” donde el premio a obtener como primer lugar es de 11’190.158 USD [5]. Existen otras ligas donde fácilmente el primer lugar puede ganar 350.000 USD [6].

En la actualidad los equipos profesionales compiten en una liga mayor llamada “The International” realizada anualmente en diferentes países e inicio en el 2011 en Colonia Alemania. Esta competencia tiene un modo para los usuarios “amateur” llamado el compendio donde aquellos jugadores que compran un pase de batalla en la temporada de la liga tienen derecho a votar por su equipo favorito. Esta característica tiene la posibilidad de completar en un árbol de posibles equipos ganadores hasta predecir qué equipo puede ser el ganador de esta liga [7].

El juego en este punto es muy versátil, cada liga tiene unos brackets muy particulares, es decir, las fases son muy parecidas a un mundial de Futbol, tiene clasificados los mejores equipos, y su apertura inicia con una nueva combinación de equipos para cada nueva versión de la liga.



[8]

Se realizan diferentes tipos de ligas en Dota 2: ligas premier, ligas mayores y ligas menores y algunas líneas amateurs que son online. Al año, se juegan alrededor de 10 ligas premier [9], 17 ligas mayores [10], cada una de estas partidas se puede asistir presencialmente o ser espectador por algún canal de streaming como twitch.tv para ver cada partida. En un día, esta transmisión puede llegar a tener un pico de 14 millones de espectadores simultáneos observando este tipo de eventos [11].

Machine Learning

Machine learning es un amplio campo de las ciencias de la computación que consiste en diseñar eficientemente algoritmos capaces de predecir estados de objetos abstraídos de la realidad [12]. También exige tener una noción de conocimiento sobre el conjunto de datos a ser estudiados, pues es importante tener cierta noción de balanza para encontrar las variables que mejor se ajusten la exactitud de la predicción.

Este campo es basado en la combinación de conceptos fundamentales de las ciencias de la computación con ideas traída de la estadística, probabilidad y optimización.

Algoritmos de aprendizaje tienen un largo camino satisfactorio desplegado en una variedad de aplicaciones en el mundo real como son [12]

- Clasificación de texto o documentos (detección de spam)
- Procesamiento de lenguaje natural (análisis morfológico)
- Reconocimiento de voz (verificación del hablador)
- Reconocimiento de un perfil óptico (OCR)
- Aplicaciones en computación biológica (funciones de proteínas)
- Tareas de computación de visión (reconocimiento de imágenes)
- Detección de fraude (detección de un intruso a una red, tarjetas de crédito)
- Juegos (Jugadores de ajedrez)
- Control de vehículos autónomos (navegación de carros autónomos)
- Diagnostico medico
- Sistemas de recomendacion

Los modelos son construidos para emparejar las relaciones desde los atributos a la clases a predecir. Si este atributo es un valor nominal, es llamada la clase y el modelo un clasificador, allí es donde hay bastante variedad de algoritmos de aprendizaje debido a los diferentes modelos de operación, para construir modelos que funcionan en su interior de manera diferente, es decir cada uno tiene su propia implementación si lo vemos desde el punto de vista estratégico.

METODOLOGÍA

La metodología que ya se encuentra en progreso en este proyecto de investigación se presenta a continuación, seguida de su significado en cada etapa.

Adquisición de datos:

Diferentes autores han utilizado el API de Valve de manera directa [14], sin embargo, después de realizar una búsqueda en tal investigación se concluye que los repositorios ya no se encuentran en funcionamiento o están fuera de mantenimiento por lo cual es necesario consulta opendota.com.

En esta etapa se presenta el API de opendota.com y se construye un software capaz de obtener los datos por cada jugador profesional que activamente pertenezca a un equipo de una liga profesional. Este proceso se encuentra en progreso, en la actualidad para la construcción del microservicio se eligió Spring Boot por su facilidad y adaptabilidad a llamados REST, almacenando todos estos llamados en una Base de datos mongoDB.

Los criterios de selección de variables eran desconocidos desde un principio, por lo cual se optó por almacenar completamente en la Base de Datos el registro como un string. Esto nos permite posteriormente poder acceder completamente a los datos.

Sin embargo, el contra de esta elección es el esfuerzo de software de la maquina que procesará estos registros para luego obtener el compendio de datos, este compendio deberá ser dividido por cada uno de los jugadores profesionales.

Tratamiento datos: Esta etapa se trata de garantizar la homogenización de datos encontrados a lo largo de los documentos json, dependiendo de lo que requiera los diferentes tipos de algoritmos que se pretendan implementar.

Problemas encontrados: Existe una creciente preocupación en el tratamiento de los datos en cuanto se ejecuta la recolección de los datos, debido al alto procesamiento de datos, obtener más de 160.000 registros paginados traídos de la base de datos, luego por cada uno de estos registros, abstraer los campos interesados, convertirlos en un objeto plano de Java para luego ser escrito en un archivo CSV en disco, hace que 16 GB de Ram sean muy poco.

Una solución permanente para este problema es crear una nueva colección en la base de datos únicamente conformada por las variables que después de los experimentos queramos mantener y hacer reducir el procesamiento. Esta solución daría como resultado asegurar la continua descarga de información del API de open dota y evitar que esto se convierta en un riesgo por limitaciones de Hardware.

Experimentos con algoritmos de aprendizaje supervisado: En esta etapa se pretende ejecutar diferentes algoritmos con un data set inicial, para determinar que algoritmo se puede ejecutar los data set finales.

Primeros resultados y conclusiones: Los resultados del análisis de los experimentos realizados en la anterior etapa, servirán no solamente para determinar que maquina se ajustaran mejor a esta investigación, sino también para tener un mejor entendimiento acerca de los valores resultantes como lo es accuracy, sensibilidad y especificidad del algoritmo elegido.

Compendio de datos para jugadores profesionales: Se pretende generar un algoritmo, que colecciones los datos de la Base de Datos, generando diferentes CSV para cada jugador y almacenarlos en un subversionador.

Implementacion Python usando sklearn - ANN: Esta etapa es meramente implementar el modelo predictivo que nos llevará a los resultados, cargando los datos de la anterior etapa usando Python y una de sus librerías (sklearn) más significativas y maduras para machine learning.

Análisis Primeros resultados jugadores profesionales: Esta etapa se encarga de analizar los resultados obtenidos, de ser necesario, se tendrá que refinar los argumentos de los modelos predictivos en python.

Compendio de datos para equipos profesionales y ligas: Se pretende almacenar y agrupar en diferentes archivos csv, los datos por liga y por equipos

profesionales que participan en las ligas. Estos datos vienen del mismo API open dota, y tendrán que ser tratados de una forma diferente a la que se implementó para jugadores singulares.

Implementación equipos profesionales en Python usando sklearn – RF: Esta etapa se encargará de implementar en phyton un modelo predictivo aplicando Random Forest usando la librería sklearn.

Resultados equipos profesionales: Determinar los resultados después de la ejecución del modelo predictivo, analizando que sus valores de accuracy sean confiables.

Conclusiones: Los resultados de los estudios y analisis previos serán suficientes para servir como pruebas relevantes de nuestros objetivos dando por terminado la metodología y que sirvan como inspiración para otros proyectos investigativos en las áreas de los e-sports.

Objetivos

Los objetivos son la pieza clave, en esta sección presento el objetivo general y los objetivos específicos.

Objetivo General

Diseñar un modelo predictivo usando máquinas de aprendizaje supervisado que prediga el equipo ganador en un torneo o liga de Dota 2 dado los datos parciales recopilados de ligas anteriores y partidas comunes de cada jugador profesional que participe en estas ligas.

Objetivos Específicos

- Determinar cuál es la relación entre las partidas que juega un jugador profesional fuera de la liga y el rendimiento de ese jugador cuando participa en una liga profesional.
- Proponer un modelo predictivo de que un equipo pueda o no ganar una partida en una liga teniendo en cuenta el historial de ligas.
- Elaborar un modelo en el cual se analicen los resultados singulares de los jugadores profesionales en sus partidas fuera de campeonato.

Preguntas de investigación

P1: ¿Cuál es la precisión de nuestro modelo predictivo para determinar si un equipo puede o no ganar una partida en una liga?

P2: ¿Cuál es la relación entre las partidas singulares y el rendimiento de un jugador en medio de una liga?

P3: ¿Cuál es la precisión más alta alcanzable usando diferente atributos en la ejecución de Random Forest?

Avances

Obtención de los datos

En orden de traer las partidas de los jugadores profesionales, fue necesario usar el API de opendota.com y hacer diferentes llamados al API Rest. Estas partidas son guardadas una vez son finalizadas, no se trata de información en vivo.

Durante la descarga de estos datos, es necesario hacer una pausa por cada llamado al API, lo cual hace que obtener todo el data set sea más lento.

A la fecha, se han descargado aproximadamente 160.000 registros de diferentes jugadores profesionales, uno de los principales retos, se encuentra en almacenar todos los datos, debido a que en un inicio no se tenía idea de cuales serían las variables que son relevantes a la investigación, entonces se decidió almacenar todo el documento json en una columna de la Base de datos.

Sin embargo, se han hecho una serie de experimentos, logrando hacer un parsing de datos para iniciar con las primeras pruebas, la siguiente tabla muestra estas variables, que son de tipo numérico.

Variables Independientes

Assists	Hero ID	Hero Kills	Lane
Neutral Kills	Game Mode	Total XP	Total Gold

Variable dependiente

'Win', es la variable o clase que queremos predecir que será de tipo numero 1 o 0. Que representa si ese jugador gana o no la partida.

Experimentos Preliminares

Se hizo un muestreo inicial para determinar cual seria el mejor algoritmo de acuerdo con su accuracy, teniendo en cuenta las tablas de falsos positivos. La

siguiente tabla muestra un breve resumen de lo obtenido con la evaluación comparativa.

Algoritmo	Accuracy	Sensibilidad	Especificidad
Naive Bayes	71.12	0.72	0.69
ANN	77.36	0.84	0.67
KNN K=1	58.85	0.63	0.52
KNN K=7	64.08	0.71	0.54
SVM	62.52	0.88	0.29
Muestra total 10.000 partidas de jugadores profesionales			

Estos clasificadores fueron probados usando 10.000 partidas de un jugador, aquí podemos ver que el algoritmo con mejores resultados resulta siendo Artificial Neural Networks, donde dado su accuracy, además de su sensibilidad especificidad.

¿Por que elegir ANN?

En este caso la sensibilidad, que nos muestra la capacidad de detectar aquellas partidas que en efecto ganaron por el jugador y la especificidad aquella que nos proporciona la ausencia de partidas no ganadas, en aquellas partidas que efectivamente ganó.

Experimentos con mas jugadores

Para la realización de este experimento, se ha adoptado por descargar la información recolectada a lo largo de este tiempo, y hacer el parsing de la columna JSON, para construir diferentes archivos CSV y hacer el debido procesamiento.

Se decidió que el mejor lenguaje para hacer el tratamiento de los datos descargados es Python, ya que este dispone de librerías domo panda para tratamiento de archivos csv, numpy para tratamiento de datasets, sklearn para hacer uso de algunos algoritmos como ANN.

Con relación a diferentes jugadores profesionales, a continuación se encuentra la primera tabla con los resultados, tomando 67% como aprendizaje y 33% el restante como test data. Por otro lado, 3 capas con 10 nodos cada uno como argumentos del clasificador. Los resultados en la siguiente tabla:

playerID	Accuracy	playerID	Accuracy
105248644	0.63	34505203	0.51
72312627	0.68	82262664	0.74
101356886	0.70	106573901	0.58

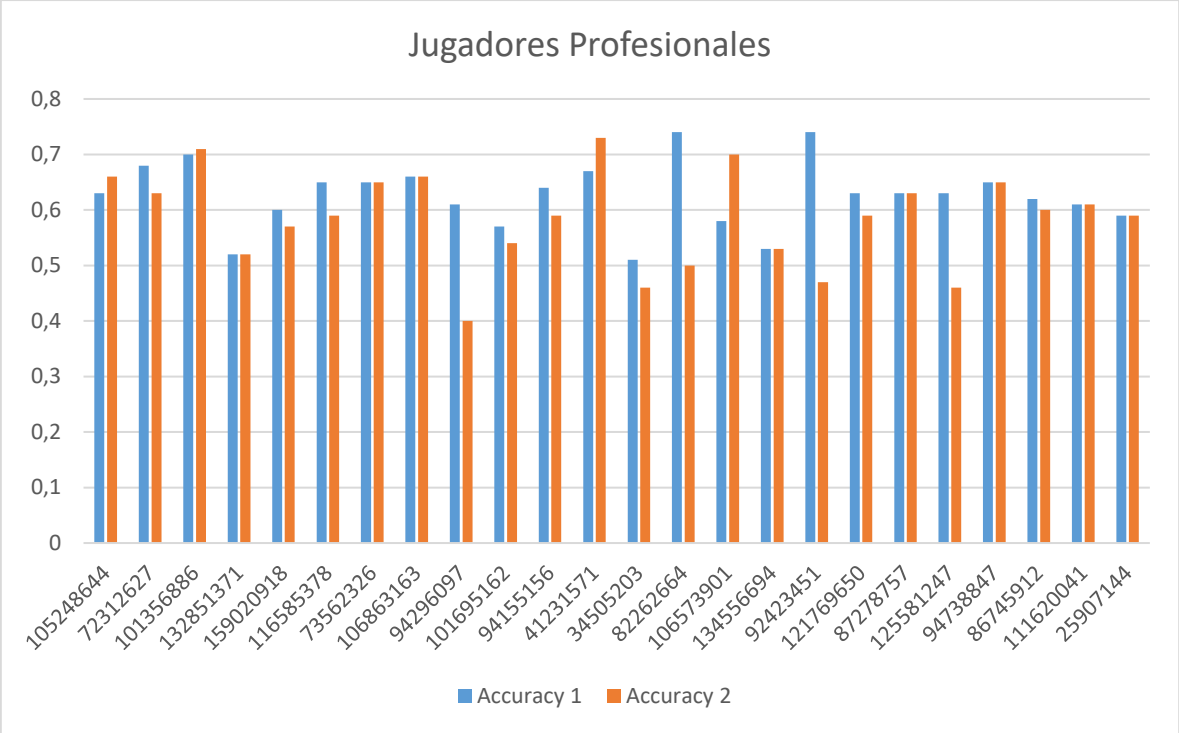
132851371	0.52	134556694	0.53
159020918	0.60	92423451	0.74
116585378	0.65	121769650	0.63
73562326	0.65	87278757	0.63
106863163	0.66	125581247	0.63
94296097	0.61	94738847	0.65
101695162	0.57	86745912	0.62
94155156	0.64	111620041	0.61
41231571	0.67	25907144	0.59

Para comparar los resultados, se decidió cambiar el data set de datos de aprendizaje, subiendo a 80% de los datos de aprendizaje, y disminuir a 2 capas y 10 nodos cada uno, los resultado en la siguiente tabla.

playerID	Accuracy	playerID	Accuracy
105248644	0.66	34505203	0.46
72312627	0.63	82262664	0.50
101356886	0.71	106573901	0.70
132851371	0.52	134556694	0.53
159020918	0.57	92423451	0.47
116585378	0.59	121769650	0.59
73562326	0.65	87278757	0.63
106863163	0.66	125581247	0.46
94296097	0.40	94738847	0.65
101695162	0.54	86745912	0.60
94155156	0.59	111620041	0.61
41231571	0.73	25907144	0.59

Si comparamos los valores obtenidos, nos damos cuenta que nuestra primer tabla se comporta para la mayoria de casos mucho mejor donde solo tenemos 2 capas y elejimos 80% de entrenamiento.

Esto debido a Overfitting, ocurre cuando un modelo aprende los detalles y el ruido en los datos de entrenamiento en la medida en que impacta negativamente el rendimiento del modelo en datos nuevos. Esto significa que el ruido o las fluctuaciones aleatorias en los datos de entrenamiento son recogidos y aprendidos como conceptos por el modelo. El problema es que estos conceptos no se aplican a nuevos datos y tienen un impacto negativo en la capacidad de los modelos para generalizar.



TRABAJO REALIZADO – ESTADO DEL ARTE

On using Artificial Neural Network models to predict game outcomes in Dota 2

Esta tesis se implementa redes neuronales y exploran con este método la selección de personajes o héroes del juego, este modelo considera únicamente dicha selección y hace una variación de neuronas para analizar su comportamiento a medida que aumentan las capas, dando a entender que es proporcional al número de predicción llegando a un 59% [13].

Result Prediction by Mining Replays in Dota 2

Esta tesis, idea un modelo de aprendizaje basado en el estado en diferentes tiempos durante un juego, predice un equipo ganador dado estos diferentes estados, implementando Random Forest después de hacer pruebas con otros algoritmos concluyendo porqué RF fue el mejor candidato para predecir la victoria sobre una partida [14].

Outcome Prediction of DOTA2 Based on Naïve Bayes Classifier

Aunque los supuestos independientes son a menudo inexactos, de hecho, algunas de las propiedades del clasificador Naive Bayes lo hacen sorprendentemente efectivo en la práctica. El autor proporciona una forma de analizar las alineaciones y la probabilidad de ganar en el Dota2 con el clasificador Naive Bayes, presenta la idea básica de cómo analizar el juego del clasificador Naive Bayes y verifica la posibilidad de analizar el juego con datos cuantitativos en el modelo de clasificador Naive Bayes [15].

Predicting Future States in DotA 2 using Value-split Models of Time Series Attribute Data

En este trabajo, se introduce un enfoque para pronosticar cambios en la salud del héroe en DotA 2 al dividir los datos en cambios grandes y pequeños, usando este enfoque de división de valores, se predice ambos tipos de cambios por separado utilizando diferentes modelos estadísticos. Para cambios pequeños, usan un modelo de media móvil autorregresiva (ARMA) y para cambios grandes usan una combinación de métodos. Los cambios grandes ("puntos de salto") se predijeron utilizando una estimación no homogénea del proceso de puntos de Poisson, mientras que la regresión logística y la regresión lineal se usaron para predecir el signo y la magnitud de estos puntos, respectivamente [16].

Prediction of Dota 2 Game Result

La parte teórica de esta tesis se centra en aclarar brevemente el árbol de decisiones y la teoría de redes neuronales artificiales, explica los factores básicos que tienen un impacto significativo en el resultado del juego. En la parte práctica, el enfoque se centra en experimentar con los parámetros de la técnica de aprendizaje automático, extender los datos de entrada con información sobre las composiciones de héroes, comparar y evaluar el rendimiento de estas extensiones. Todo esto da como resultado la implementación de un programa experimental que producirá un modelo ANN predictivo. Este modelo se puede usar más adelante para predecir el resultado del juego según el conocimiento de las composiciones del héroe del equipo inicial [17].

Bibliography

[1] M. G. Wagner, "On the Scientific Relevance of eSports," in *Proceedings of the 2006 International Conference on Internet Computing & Conference on Computer Games Development*, Las Vegas, Nevada, USA, 2006.

[2] K. Noonan, "The Motley Fool: Stock Investing Advice | Stock Research," The Motley Fool, 26 11 2017. [Online]. Available: https://www.fool.com/investing/2017/10/25/7-gaming-stats-that-prove-esports-is-the-next-big.aspx. [Accessed 30 11 2018].

[3] esc.watch, "Researching esports and streaming trends," 27 08 2018. [Online]. Available: https://esc.watch/storage/app/media/uploaded-files/TI_All_Stages.png. [Accessed 04 12 2018].

[4] Research Gate, "Discover scientific knowledge, and make your research visible.," researchgate.net, 2018. [Online]. Available: https://www.researchgate.net/figure/Map-of-Dota-2-from-Dota-2-wiki-7_fig1_262207918. [Accessed 21 11].

[5] V. Corp, "Dota 2 - The International," Valve, 20 08 2018. [Online]. Available: http://www.dota2.com/international/overview/. [Accessed 06 11 2018].

[6] Liquid, "Liquipedia Kuala Lumpur Major," Team Liquid, 2018. [Online]. Available: https://liquipedia.net/dota2/PGL/Kuala_Lumpur_Major. [Accessed 06 11 2018].

[7] Dota2, "Battlepass TI," Dota 2, 2018. [Online]. Available: https://www.dota2.com/international/battlepass/. [Accessed 08 11 2018].

[8] Liquipedia, "Liquipedia," Liquipedia, 2018. [Online]. Available: https://liquipedia.net/dota2/The_International/2018. [Accessed 20 11 2018].

[9] T. Liquid, "Premier tournaments," Liquid, 2018. [Online]. Available: https://liquipedia.net/dota2/Premier_Tournaments. [Accessed 05 11 2018].

[10] t. Liquid, "Major tournaments," Team Liquid, 2018. [Online]. Available: https://liquipedia.net/dota2/Major_Tournaments. [Accessed 05 11 2018].

[11] "Researching esports and streaming trends," ESM.one, 2018. [Online]. Available: https://esc.watch/blog/post/stats-international-2018. [Accessed 05 11 2018].

[12] M. M, R. A and T. and A, Foundations of Machine learning, London, England: The MIT Press, 2012.

[13] W. VIKTOR and A. JULIEN, On using Artificial Neural Network models to predict game outcomes in Dota 2, Stockholm, Sweden, 2017.

[14] J. W. Filip Johansson, Result Prediction by Mining Replays in Dota 2, Karlskrona, 2015.

[15] K. Wang and W. Shang, Outcome Prediction of DOTA2 Based on Naive Bayes Classifier, Beijing, 2017.

[16] Z. Cleghern, O. Ozaltın, S. Lahiri and D. Roberts, Predicting Future States in DotA 2 using Value-split Models of Time Series Attribute Data, North Carolina, 2017.

[17] F. Beskyd, Result, Prediction of Dota 2 Game, Prague, 2018.