UNIVERSIDAD DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA QUÍMICA, BIOTECNOLOGÍA Y MATERIALES

Smart Training, Servicio web de Entrenamiento de Modelos

DAVID ALFREDO MEDINA ORTIZ

Profesor Tutor: ÁLVARO OLIVERA NAPPA, PH.D.

Resumen Desarrollo de Proyecto,
Seminario 02

Santiago – Chile
Agosto, 2018

TABLA DE CONTENIDOS

página

ÍNDICE DE FIGURAS

página

ÍNDICE DE TABLAS

página

RESUMEN

Conceptos como minería de datos, machine learning, big data, análisis estadísticos, modelamientos matemáticos, etc, son mencionados día a día, ya sea en el ámbito privado como público, involucrando áreas como: comercio, salud, investigación, transporte, etc. Lo cual denota que son temáticas que han adquirido mayor relevancia y su ascenso seguirá con el pasar del tiempo.

La manipulación de grandes volúmenes de datos, con el fin de poder extraer información de ellos, búsqueda de patrones, evaluaciones estadísticas, etc. Implica por parte del interesado, tener conocimientos en dichas áreas además de comprender herramientas informáticas que le permitan dicho procedimiento. Sin embargo, dichas herramientas o son costosas, debido a la licencia que implica, o, se requiere de conocimiento informático para su manipulación, debido a que requiere implementar módulos o servicios a medida que permitan ejecutar las tareas de interés, lo cual deja a un número importante de entidades que desean involucrarse en dicho mundo, pero no cuentan con las capacidades ni tampoco con las competencias para ello.

Dado a lo anterior y en base a la creciente demanda de desarrollo de metodologías que permitan aplicar data mining a procesos de datos, con el fin de extraer información y conocimiento de la misma, se propone Smart Training, como sistema web, que facilite los procesos de evaluaciones estadísticas, búsqueda de patrones de comportamiento, desarrollo de modelos de clasificación y evaluación de características o features en el set de datos a estudiar.

1. INTRODUCCIÓN

1.1. Data Mining

Minería de datos es el proceso de descubrimiento de patrones en set de datos, involucrando métodos asociados a Machine Learning, Estadísticas y sistemas de bases de datos. [1]. La minería de datos es un subcampo interdisciplinario de la informática, el cual tiene por objetivo general extraer información (a través de métodos inteligentes) de un conjunto de datos y transformar la información en una estructura comprensible para su uso posterior. [2, 3]. La minería de datos es el paso de análisis del proceso de *descubrimiento de conocimiento en bases de datos*, o KDD. [4]. Además del análisis en bruto de los datos, también incluye aspectos de manipulación de bases de datos, pre procesamiento de datos, evaluaciones de modelo e inferencia, métricas de interés, consideraciones de complejidad, post procesamiento de estructuras descubiertas, visualización y actualización de la información.

En la Figura 1.1, se exponen las principales ramas que componen la minería de datos y los diferentes procesos que se asocian a dichas ramas.

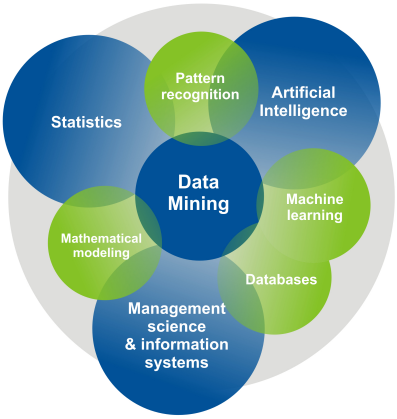

Figura 1.1: Componentes en la minería de datos

Son tres las principales áreas que abarca la minería de datos: Estadística, Inteligencia Artificial y Manipulación de sistemas de información, mientras que son distintos procesos los que interactúan entre estas ramas, tales como: Modelamiento Matemático, reconocimiento de patrones, Sistemas de almacenamiento persistente y machine learning.

Cada área en particular tiene un objetivo general y diversos objetivos específicos. Sin embargo, estas áreas interactúan entre sí, con el fin de poder extraer patrones de información que generen conocimientos a partir de la data de procesada.

La minería de datos se utiliza en diferentes campos, tales como: Genética, Evaluaciones proteómicas, Comercio, Sistemas de tránsito, Optimizaciones en procesos industriales, reconocimiento de patrones y rasgos cuantificables en enfermedades y más recientemente en áreas de dinámicas moleculares y parámetros para la generación de pipe lines automatizados de simulaciones cuánticas en sistemas químicos.

El uso de la minería de datos y la búsqueda de patrones de comportamientos de datos de interés, no sólo es un área que se enfoca en la investigación. Diversas son las entidades privadas que ofrecen servicios de big data y data science, además del sector público, con el fin de detectar puntos críticos de zonas de riesgo, zonas de accidentes, evaluación o perfilación de grupos de estudio, etc.

1.2. Problemática

Actualmente, los sistemas de almacenamiento persistente, permiten disponer de información que es de interés para distintos tipos de entidades, las cuales dependerán del área a las que se dediquen. Sin embargo, cada área en particular, tiene objetivos similares, tales como:

- Qué significa los datos que tengo?
- Puedo personalizar y interpretar datos?
- Puedo optimizar procesos en base a la información que tengo sobre estos?
- Es factible aumentar la experiencia de usuario en cuanto a procesos de ventas y compras, personalizando sus áreas de cliente?

Son muchos los objetivos que se pueden encontrar y muchas las aplicaciones que implica esta técnica. No obstante, el hecho de tener data almacenada y no saber cómo procesarla es un gran problema para muchas pequeñas y medianas empresas, así como para también ventas, laboratorios de investigación, etc.

El deseo de aplicar minería de datos, con el fin de extraer información, es día a día, más frecuente. Sin embargo, un usuario común debe enfrentar la problemática de como afrontar el problema, adquirir las competencias, o simplemente, contratar servicios de data science, los cuales cobran altas sumas de dinero y emplean un tiempo elevado con el fin de llegar a una respuesta pronta.

Por otro lado, existen herramientas que facilitan la utilización de minería de datos, pero, el costo por conceptos de licencia es demasiado elevado y no permite abarcar diversas áreas y testear algoritmos variados. Mientras que por otro lado, existen módulos o librerías que han sido implementadas en diversos lenguajes de programación, con el fin de aplicar minerías de datos, pero esto aumente aún más la complejidad del tema, debido a que para implementar dicha labor, se requiere de conocimientos en programación y en algunos casos, las implementaciones son engorrosas y requieren de un conjunto de arquitecturas que soporten dichas instancias.

Dado a lo anterior, es que se propone el desarrollo de una herramienta web, que permita la aplicación de diversas técnicas de minería de datos y oriente de manera inteligente al usuario, esto implicaría que no se requiere de un conocimiento de programación, además que las competencias en minería de datos deben ser mínimas, puesto que la idea contempla la orientación al usuario con respecto al objetivo que desea.

Durante este documento, se expone el diseño de la herramienta, las metodologías a utilizar y los artefactos de software que se crearán, con el fin de poder implementar esta herramienta en base a una metodología de diseño, además se expone un resumen de las tecnologías actuales, cuales son las ventajas y desventajas que poseen cada una y en qué se diferencia el software planteado con respecto a los existentes.

1.2.1. Estado del Arte

1.3. Smart Training

Smart Training es un sistema web, que facilita la utilización de técnicas de minería de datos, con el fin de evaluar características en set de datos, reconocer patrones, entrenar algoritmos de clasificación, generar evaluación de características, etc. Tiene por finalidad acercar la minería de datos a un público que no posee las competencias para implementar modelos mediante utilización de módulos de programación.

Smart Training se compone de 5 módulos, los cuales se detallan a continuación.

1.3.1. Módulo de Procesamiento de Datos

Este módulo tiene por objetivo cargar la data entregada en archivos de texto, evalúa los datos existentes, corrobora que no existan problemas, revisa el set de datos, encuentra valores nulos, codificaciones no permitidas, etc, a lo que, finalmente, entrega un resumen del proceso, exponiendo los resultados y de dicha tarea y comentando si es posible trabajar con dicho set cargado, en caso contrario, expone mensajes con recomendaciones a seguir.

1.3.2. Módulo de Análisis Estadístico

Este módulo permite la evaluación del set de datos contemplando, correlaciones, box plot, distribuciones mediante histogramas, scatter plot, gráficos de densidad de datos, además de resúmenes estadísticos para cada atributo o característica en el set de datos que se está trabajando.

1.3.3. Módulo de Análisis de Features

Este módulo permite evaluar las características en el set de datos y el aporte que éstas entregan, adicional a ello, contempla análisis mediante técnicas PCA, para generar reducciones de dimensionalidad, Mutual Information y técnicas basadas en correlación, todas con el objetivo de explicar los comportamientos de éstas y cómo influyen en el set de datos.

1.3.4. Módulo de Clustering

Clustering es una de las técnicas más conocidas para asociar segmentos en una muestra, es decir, generar grupos o particiones que tengan un alto grado de diferencia entre ellas, pero cuyos integrantes en una partición dada, sean altamente similares.

Existen diferentes algoritmos de clustering y parámetros asociados a estos, los cuales tienen formas de encontrar particiones distintas, basándose en distancias, medidas gausianas, generación de hiper planos, etc.

Este módulo tiene por objetivo generar exploración de dichas técnicas y algoritmos, con el fin de poder entregar particiones en el set de datos, además, permite la evaluación de dichas particiones con el fin de poder determinar si son estadísticamente significativas o no, además de cumplir con los criterios de similitud y diferenciación mencionados previamente.

Normalmente, la búsqueda de particiones conlleva al hecho de generar modelos de clasificación para nuevos ejemplos y determinar a qué particiones se encuentran, o también, generar divisiones para implementar set de datos diferentes con comportamientos

diferentes, de tal manera que a la hora de aplicar algoritmos de clasificación, sus comportamientos presenten mayor eficiencia.

1.3.5. Módulo de Entrenamiento de Modelos

Entrenar un modelo de clasificación, predicción, implica tener un conjunto de elementos con su clasificación o valor de predicción conocido, con el fin de poder, a partir de éste, evaluar nuevos ejemplos, ya sea para clasificarlos o para predecir posibles valores de interés. Todas estas tareas, aplicando minería de datos, se logran mediante la implementación de algoritmos de aprendizaje supervisado.

Existen diferentes algoritmos de aprendizaje supervisado, los cuales contemplan diferentes formulaciones matemáticas y características, los cuales cumplen con dicha tarea, cada uno de estos, presenta características distintas, en relación al funcionamiento del mismo, la elección de un algoritmo por sobre otro, va de la mano con el hecho de las necesidades que el problema conlleva, ya sea, con el fin de entregar sólo un resultado, adicionar valores que permitan explicar el porqué de la clasificación, etc. Normalmente, para un problema desconocido, es necesario implementar fases exploratorias que permitan evaluar diferentes algoritmos y sus parámetros. Con el fin de poder, a partir de dicha instancia y en base a métricas que permitan evaluar el desempeño, seleccionar un algoritmo y sus parámetros.

Lo anterior, es el objetivo del módulo de entrenamiento de modelos, la idea de éste, es recibir un set de datos con ejemplos clasificados o cuya respuesta tenga un valor conocido y aplicarle diversos algoritmos y variaciones de parámetros, entregando un resumen de las medidas de desempeño obtenidas y efectuando un ranking según medida, para que finalmente se pueda entregar una recomendación de los mejores modelos para un cierto problema.

1.3.6. Diagrama solución

1.4. Objetivos y Alcances

El objetivo general del proyecto contempla la implementación de un sistema web denominado, *Smart Training*, el cual permita la aplicación de distintas técnicas de minería de datos a set de datos de interés del usuario, a partir de los cuales éste pueda entender comportamientos de datos y generar conocimiento a partir de ellos.

Es importante destacar los objetivos específicos que nacen dentro del desarrollo del proyecto.

1. Diseñar metodología de software a utilizar.
2. Comprender los requerimientos observados a partir del estado del arte.
3. Evaluar las funcionalidades y atributos que tendrá el sistema.
4. Comprender los usuarios y actores del software.
5. Entender las secuencias y flujos de trabajo existentes en la herramienta.
6. Crear modelos de conceptos, entidades y clases.
7. Implementar los módulos propuestos.
8. Implantar sistema.

1.5. Metodología de Desarrollo de Software

Existen diversas metodologías de desarrollo de software, las cuales contemplan diferentes características y se enfocan en distintos puntos objetivo. Algunos ejemplos son.

- Metodologías Ágiles.
- Diseño cascada.
- Diseño estreslla.
- Iterativas.

Metodología ágil, se utiliza cuando el objetivo se basa principalmente en sacar a producción el software de manera rápida, no contempla procesos de diseño complejos y simplemente se centra en el desarrollo del producto, lo cual permite, por un lado, poseer un producto en poco tiempo. Sin embargo, está sometida a enmarcar errores debido a que no se consideraron patrones de diseño.

El diseño en cascada y también en estrella, se centran en los requerimientos de usuario y generar sub productos asociados al software final, los cuales cumplen un objetivo en particular del software, fueron muy utilizadas en los años 90, debido a la simplicidad que estos poseían.No obstante, no contempla iteraciones para evaluar los flujos de trabajo, ni tampoco la utilización de paradigmas complejos de desarrollo de software.

Una de las metodologías más utilizadas, es la Iterativa, ésta contempla un conjunto de patrones de diseño, los cuales están asociados a la evaluación de las funcionalidades, los atributos, describir secuencias de flujos y asociar conceptos, es la metodología que

contempla un mayor conjunto de pasos. No obstante, es la que más asegura que a la hora de implementar, dicho proceso sea rápido. El hecho de ser iterativa, implica que cada etapa entrega un artefacto de software, bajo el cual depende el siguiente, en nuevas etapas se hacen mejoras a los artefactos generados y se está en constante feedback.

Adicional a las metodologías de software, existen diferentes paradigmas de programación que son empleados en conjunto con dichas estrategías. Los dos principales son: Estructurado y Orientado a Objetos. En el primero, se sigue un orden secuencial del problema a desarrollar, no está adaptado para grandes desarrollos, si no que más bien, es empleado en scripting y manejo de patrones en archivos de texto. Por otro lado, la programación Orientada a Objetos (POO), cumple con la características de ser más cercana a la vida real, debido a que se basa en el diseño de clases, que representan entidades las cuales pueden ser ficticias o reales. Presenta grandes ventajas debido a que posee las siguientes características.

- *Encapsulamiento*: Un Objeto es dueño de sus atributos y métodos.
- *Polimorfismo*: Un mismo método, pueden tener significados diferentes para distintas clases.
- *Reusabilidad*: Una clase modelada puede utilizarse en diversos proyectos debido a que siempre poseerá los mismos atributos y métodos.

Además de dichas características, la POO asocia conceptos abstractos a la programación, tales como: herencia, dependencias, asociaciones y composiciones, las cuales aumentan más aún, la usabilidad de este paradigma.

En esta ocasión, debido a la envergadura del proyecto, a las características que se espera que posea y a las ventajas que entregan las metodologías, se utilizará el diseño de software iterativo con objeto acoplado. Es decir, se diseñará teniendo en consideración distintas iteraciones asociadas a los artefactos de software que se desarrollen, enfocando dicho diseño a POO.

2. Análisis

La etapa de análisis del proyecto, contempla la evaluación y determinación de las diferentes funcionalidades que tendrá el sistema, asociado a los atributos que estos presentan y que permiten cuantificar de cierta manera el software. También contempla las evaluaciones de los flujos asociados a cada función y determina las secuencias de pasos a seguir para dar respuesta a cada una de éstas, exponiéndolos en forma narrativa mediante los casos de uso. Finalmente se evalúan los conceptos existentes y que representan entidades dentro del software, los cuales forman parte del diseño posterior.

2.1. Funciones del Sistema

2.2. Atributos del Sistema

2.3. Actores y Usuarios

2.4. Casos de Uso

2.4.1. Diagramas de casos de uso

2.5. Diagramas de secuencia o colaboración

2.6. Conceptos

2.6.1. Modelo Conceptual

2.7. Entidades

2.7.1. Modelo de Entidades

3. Diseño

3.1. Arquitectura de Software

3.2. Diagramas de Interacción

3.3. Diagrama de Clases

3.4. Diagramas de Estado

4. Planificación

4.1. Etapas del Proyecto

Bibliografía

[1] *Data Mining Curriculum*. ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[2] Clifton, Christopher (2010). *Encyclopedia Britannica: Definition of Data Mining*. Retrieved 2010-12-09.

[3] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Retrieved 2012-08-07.

[4] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). *From Data Mining to Knowledge Discovery in Databases* (PDF). Retrieved 17 December 2008.