

Universidad Nacional de Asunción

Facultad Politécnica

Ingeniería Informática

Trabajo Final De Grado

TPDBM: Técnica de previsión de la demanda basada en Business Intelligence y Machine Learning

Junio 2017

1. **SÍNTESIS DEL PROYECTO**
   1. TÍTULO INICIAL DEL PROYECTO

|  |
| --- |
| TPDBM: Técnica de previsión de la demanda basada en Business Intelligence y Machine Learning |

SINTETIZAR EL PROYECTO EN TRES PALABRAS CLAVE

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Previsión de la Demanda |  | Business Intelligence |  | Machine Learning |

* 1. DURACIÓN DEL PROYECTO

|  |  |
| --- | --- |
| 12 | **MESES** |

* 1. DATOS DE LOS ESTUDIANTES Y DEL ASESOR DEL PROYECTO

|  |  |
| --- | --- |
| **NOMBRE DEL ASESOR PRINCIPAL:**  Ph.D. Diego Pinto | |
| **Email:** dppintoroa@gmail.com | **Teléfono:** 0984 401 410 |
| **CARGO:**  Docente | **FIRMA:** |

|  |  |
| --- | --- |
| **NOMBRE DEL ASESOR SECUNDARIO:**  Ing. Aditardo Vázquez | |
| **Email:** vazquez.aditardo@gmail.com | **Teléfono:** 0981 358 294 |
| **CARGO:**  Docente | **FIRMA:** |

|  |  |
| --- | --- |
| **NOMBRE DEL ALUMNO:**  Alberto Gabriel Garcete González | |
| **CI:** 3.676.206 | **PLAN DE ESTUDIOS:** 2008 |
| **Email:** albertogarcetepy@gmail.com | **Teléfono:** 0971 494427 |
| **AÑO DE INGRESO: 2008** | **FIRMA:** |

|  |  |
| --- | --- |
| **NOMBRE DEL ALUMNO:**  Raúl Alberto Benítez Martínez | |
| **CI:** 4.089.446 | **PLAN DE ESTUDIOS:** 2001 |
| **Email:** raulkvd@gmail.com | **Teléfono:** 0984 196134 |
| **AÑO DE INGRESO: 2007** | **FIRMA:** |

* 1. **RESUMEN DEL PROYECTO**

La previsión de la demanda supone uno de los procesos más importantes y de mayor impacto económico en empresas relacionadas a las ventas minoristas. Dentro de la gestión de compras, un cálculo que regularmente debe hacerse es este pronóstico de la demanda mencionado, para estimar el volumen de productos de la orden de compra y con esto realizar la reposición de stock. En este trabajo se propone una técnica que ayuda a tomar decisiones acerca del volumen de la orden de compra para la reposición de stock. Esta técnica está basada en Business Intelligence y Machine Learning integradas, se trata de un modelo que pronostica la demanda de productos para el siguiente período de ventas, a su vez que también incluye decisiones de si realmente hay que seguir adquiriendo un determinado producto. Los experimentos indican que se puede automatizar el proceso de previsión de la demanda y las decisiones del volumen de productos en las órdenes de compra para reposición de stock. Este modelo ofrece altas tasas de aciertos.

Previsión de la Demanda

En las empresas retail o de ventas minoristas uno de los principales desafíos que afrontan es el manejo eficiente de stock. La eficiencia se busca de manera a evitar tener excesivas cantidades de productos en los depósitos, o evitar la falta de productos denominada ruptura de stock. Los excesos incurren en sobrecostos, lo contrario conlleva a pérdida de oportunidades de ventas al no disponer del producto en el momento oportuno. Ambos repercuten directamente en las utilidades de la empresa y en la satisfacción de los clientes. El problema a resolver consiste en estimar o predecir la cantidad de productos a vender en el próximo período de venta.

Actualmente, en el proceso de gestión de compras para la reposición de stock se utilizan técnicas de pronósticos para determinar las cantidades de las órdenes de compra. Estas técnicas pueden estar basadas en pronósticos cuantitativos o cualitativos. Los modelos de cantidad fija y los modelos de período fijo son ampliamente utilizados. Independientemente de la técnica elegida el problema real con los pronósticos es su falta de confiabilidad, ya que por lo general no son precisos. La interrogante que siempre surge en estos modelos es si sus resultados serán superiores o inferiores a la demanda real y en qué medida.

Business Intelligence (KPI)

Con el presente trabajo se elabora un nuevo modelo de estimación de cantidades eficientes en las órdenes de compra de productos para la reposición de stock del siguiente periodo de venta. En este nuevo modelo se integran técnicas de Business Intelligence y Machine Learning.

En la etapa de Business Intelligence el objetivo principal es calcular los Indicadores Claves de Rendimiento (KPI - Key Performance Indicators) de los productos en base a los datos históricos obtenidos de la base de datos transaccional. Luego cada serie de KPI obtenidos pasan por un proceso de etiquetado, donde el experto en compras los analiza y determina qué nivel de compra conviene para cada serie de KPI.

Machine Learning

En la etapa de Machine Learning se utilizan como entrada las series de KPI obtenidas en la etapa de Business Intelligence. Estas entradas constituyen las instancias que alimentan los distintos algoritmos de clasificación del Machine Learning supervisado. Luego tienen lugar los procesos propios de esta etapa que son el entrenamiento y testeo para finalmente evaluar los distintos desempeños a fin de determinar los algoritmos más adecuados. Los algoritmos elegidos serán utilizados para estimar las cantidades de las órdenes de compra por cada producto.

Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Los algoritmos pueden ser aplicados directamente a un conjunto de datos o llamados desde código Java. Weka contiene herramientas para pre-procesamiento de datos, clasificación, regresión, clustering, reglas de asociación y visualización. También es adecuado para desarrollar nuevos esquemas de aprendizaje automático [Weka3]. Los algoritmos de clasificación de Weka que se utilizarán son los siguientes [DM2016]: BayesNet, NaiveBayes, NaiveBayesUpdateable, Logistic, MultilayerPerceptron, SimpleLogistic, SMO, OneR, DecisionTable, JRip, PART, ZeroR, DecisionStump, J48, LMT, RandomForest, RandomTree, REPTree.

La evaluación es la clave para lograr avances reales en el aprendizaje automático. En este trabajo se aplican dos técnicas de evaluación, una de ellas es la Validación Cruzada (Cross-Validation) que consiste en dividir los datos en un número de pliegues o particiones. Si por ejemplo elegimos cuatro, entonces cada partición se utiliza para las pruebas y las demás para el entrenamiento. Al repetir este proceso 4 veces se consigue que cada partición se haya utilizado una vez como conjunto de pruebas. La técnica estándar para predecir la tasa de error es la Validación Cruzada Estratificada (Stratified k-fold Cross-Validation). La estratificación se refiere al proceso de reorganizar los datos de tal manera a asegurar que cada pliegue sea una buena representación del conjunto. Comúnmente se acepta que 10 es el número de pliegues con el que se obtiene la mejor estimación de error, idea basada en diversas pruebas sobre conjuntos de datos diferentes y para distintas técnicas de aprendizaje [DM2011].

Otra técnica es el Porcentaje de División (Percentage Split) con el que puede retener para la prueba un determinado porcentaje de los datos. Es una alternativa utilizar un conjunto de pruebas separado o una división porcentual de los datos de entrenamiento. Si elegimos 60% como porcentaje de división, entonces el conjunto de prueba se constituirá con el 40% de las instancias y el conjunto de entrenamiento con el 60% de las instancias.

Experimentos

Cada producto está asociado a un conjunto de instancias referentes a BI mensuales, a un conjunto de instancias referentes a BI quincenales y a un conjunto de instancias referentes a BI semanales. Es decir, cada producto está asociado a tres conjuntos de instancias diferentes.

Luego, cada conjunto de instancias se entrena con todos los algoritmos de clasificación WEKA posibles y la evaluación se hace tanto por el método Percentage Split así como también por el método Stratified K-fold Cross Validation. Finalmente se obtienen métricas de evaluación del aprendizaje automático con el método Percentage Split para los periodos mensuales, quincenales y semanales; y métricas de evaluación del aprendizaje automático con el método Stratified K-fold Cross Validation para los periodos mensuales, quincenales y semanales.

Por cada modelo procesado, luego de construir su clasificador y evaluarlo se obtienen las métricas Cantidad de Aciertos o el Porcentaje de Acierto, la estadística Kappa; y por cada clase (Nada, Medio, Mucho) se obtienen las métricas Area Under ROC (ROCA), Recall (RCALL), Precision (PREC), F-Measure (FMEA) y Area Under Precision-Recall Curve (PRCA).

El análisis global de los resultados se basa en la métrica Kappa. Por cada producto analizado se elige como clasificador aquel que haya alcanzado el mayor valor de Kappa, para luego calcular tres porcentajes de aciertos promedios globales: la tasa de aciertos para periodos mensuales, quincenales y semanales.

En cuanto a las limitaciones de este nuevo modelo se puede mencionar que no toma en cuenta los costes relacionados al inventario: como costes de mantenimiento, de personal, seguros, etc. El modelo planteado se aplica a empresas retail dedicadas a la venta de productos terminados.

* 1. **OBJETIVOS DEL PROYECTO.** 
     1. OBJETIVOS GENERALES
* Modificar y extender el algoritmo TKU para que pueda ser un algoritmo paralelo distribuido y comparable con otros algoritmos, permitiendo así la optimización del tiempo de descubrimiento de high utility itemsets en volúmenes de datos muy grandes.
  + 1. OBJETIVOS ESPECÍFICOS
* Implementar una modificación del algoritmo TKU para poder realizar comparaciones con otros algoritmos que asemejan su implementación en cuanto a entrada y salida del mismo.
* Extender el algoritmo TKU [3] mediante modelos como map reduce de manera que pueda ser paralelizable.
* Probar el algoritmo desarrollado y comparar los resultados con el algoritmo TKU.

1. **FORMULACIÓN GENERAL DEL PROYECTO**
   1. **PROBLEMA E HIPÓTESIS**

2.1.1 ¿CUÁL ES EL PROBLEMA U OPORTUNIDAD?

Existen muchos algoritmos para descubrir High Utility Itemsets, con distintos enfoques, estos algoritmos han ido mejorando con el tiempo enfocándose en la generación de candidatos, pero su rendimiento se ve afectado cuando se trabajan con volúmenes de datos muy grandes debido a la gran cantidad de memoria y tiempo que se necesita.

Uno de los algoritmos que presenta un buen desempeño es llamado EFIM [6] la ventaja que presenta el algoritmo de TKU sobre éste es que no requiere la entrada de un valor denominado utilidad mínima, en vez de está requiere la entrada de cuantos resultados es deseado obtener, siendo éste el valor de K.

La ventaja de requerir un valor K frente a una la utilidad mínima es que si el valor de la utilidad mínima es muy chica la cantidad de resultados será escasa y si es muy grande los resultados no serán de gran utilidad, a diferencia de si se tiene como entrada una cantidad K el resultado será más especificó proveyendo así más utilidad de resultados.

2.1.2 ¿CUÁL ES LA SOLUCIÓN PROPUESTA POR EL PROYECTO DE INVESTIGACIÓN?

Realizar la modificación del algoritmo TKU para que su comportamiento permite la comparación con otros algoritmos, en cuanto a la entrada y salida del mismo, y extender el algoritmo para buscar un mejor rendimiento mediante la paralelización y ejecución distribuida del mismo permitiendo agilizar la búsqueda de high utility itemsets mediante el framework Apache Spark?.

2.1.3 ¿CUÁL ES LA HIPÓTESIS DEL PROYECTO?

Se espera que el algoritmo PMTKU que se pretende desarrollar optimice el tiempo de búsqueda de high utility itemsets comparado con el TKU base y otros algoritmos que sigan el mismo patrón de entrada a la hora de tratar con volúmenes de datos muy grandes

* 1. **ANTECEDENTES.**

Antecedentes de Pronóstico de la demanda

Antecedentes de Business Intelligence

Antecedentes de Machine Learning

En 1959 Arthur Samuel en una publicación escribió: “Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort” [Arthur1959]. Lo que nos lleva a pensar que uno de los pioneros del aprendizaje automático ya dejaba visualizar que los programas, a partir del aprendizaje sobre los datos históricos (la experiencia), podrían efectuar tareas de toma de decisiones sin ser programadas explícitamente dichas decisiones. Samuel define al aprendizaje automático como sigue: “El aprendizaje automático es un campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas”.

Otro investigador de aprendizaje automático Tom Mitchell propuso en 1998 la siguiente definición: “Well posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E”. Donde se nos indica que el aprendizaje en las máquinas deberá ser parecido al aprendizaje en los humanos, por ejemplo cuando una criatura comienza a hablar a través de la experiencia de pronunciar las palabras y de su interacción con otras personas, entonces sucede que su capacidad de hablar se va perfeccionando o mejorando.

“The purpose of machine learning is to learn from training data in order to make as good as possible predictions on new, unseen, data” [Jean2016]. La dificultad radica en que debemos construir modelos que nos acerquen a una buan predicción sobre datos aún no conocidos o imprevistos.

Publicaciones que comprenden los últimos diez años, presentan trabajos muy interesantes en el ámbito de Machine Learning asociado al Demand Forecasting (Pronóstico de la Demanda). Se citan a continuación algunos de los problemas afrontados en publicaciones, que dan una idea del estado del arte en este tema:

* Mejorar la precisión de la previsión de demanda de agua urbana para la ciudad de Montreal – Canadá (2017).
* Proponer un método de control inteligente para sistemas de calefacción y refrigeración. (2017).
* Proponer un modelo predictivo probabilístico de consumo de energía, basado en datos, para la predicción del consumo en edificios residenciales (2017).
* Revisión de diferentes modelos de predicción de la carga eléctrica con un enfoque particular en modelos de regresión (2017).
* Aplicación de Machine Learning en la nube para encontrar conversaciones de los consumidores que influyen en las decisiones de compras (2016).
* Modelar la demanda turística de España (2016).
* Predicción del Mercado de Valores (2016).
* Análisis para un minorista en línea: Previsión de la demanda y optimización de precios (2016).
* Demanda de calefacción residencial basado en el consumo total mensual de gas natural (2015).
* Predicción de la demanda de importación de crudo en Taiwán (2014).
* Predicción del desempeño de las estrategias de pronóstico para la demanda de repuestos navales (2012).

En este trabajo, abordamos el problema de high utility itemset mining al proponer PMTKU (Parallel Modification algorithm for top-k high utility itemset). PMTKU es un algoritmo paralelo distribuido, que paraleliza a TKU y modifica la entrada del algoritmo para descubrir high utility itemset.

* 1. **METODOLOGÍAS**
* Investigación del estado del arte de:
  + High utility itemset mining
  + Parallel high utility itemset mining
* Selección del algoritmo a hacer paralelizable.
* Generación de los conjuntos de datos para las pruebas.
* Pruebas con los conjuntos de datos generados.
* Modelado del algoritmo PMTKU
* Implementación y depuración del algoritmo PMTKU
* Comparación con el algoritmo EFIM
* Redacción del documento científico
  1. **RESULTADOS ESPERADOS DEL PROYECTO**

Se espera que el algoritmo PMTKU propuesto para descubrir high utility itemsets en grandes conjuntos de datos mejore en términos de tiempo al algoritmo TKU y permita la comparación con otros algoritmos que sigan el mismo patrón de entrada.

1. **PLANEAMIENTO DEL PROYECTO**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Dic | Ene | Feb | Mar | Abr | May | Jun | Jul | Ago | Set | Oct | Nov |
| Investigación del Estado del Arte |  |  |  |  |  |  |  |  |  |  |  |  |
| Diseño del modelo propuesta. |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementación. |  |  |  |  |  |  |  |  |  |  |  |  |
| Pruebas y Ajustes |  |  |  |  |  |  |  |  |  |  |  |  |
| Escritura del Libro |  |  |  |  |  |  |  |  |  |  |  |  |
| Evaluación de Resultados y elaboración de las conclusiones |  |  |  |  |  |  |  |  |  |  |  |  |
| Defensa del trabajo |  |  |  |  |  |  |  |  |  |  |  |  |

1. **TEMARIO TENTATIVO DEL REPORTE TÉCNICO FINAL**
2. Introducción
3. Descripción del problema
4. High utility itemset mining
5. Algoritmos paralelos
6. Parallel high utility itemset mining
7. Algoritmo propuesto basado en el TKU
8. Resultados Experimentales
9. Conclusiones y trabajos futuros
10. Referencias

**REFERENCIAS BIBLIOGRÁFICAS**

[1] An introduction to frequent pattern mining. <http://data-mining.philippe-fournier-viger.com/introduction-frequent-pattern-mining/>.

[2] An introduction to High-Utility Itemset Mining. <http://data-mining.philippe-fournier-viger.com/introduction-high-utility-itemset-mining/>.

[3] Cheng Wei Wu , Bai-En Shie , Philip S. Yu , Vincent S. Tseng Mining Top-K High Utility Itemsets .

[4] Quang-Huy Duonga,, Bo Liaoa, Philippe Fournier-Viger b, Thu-Lan Dama, An efficient algorithm for mining the top-k high utility itemsets, using novel threshold raising and pruning strategies

[5] Jerry Chun-Wei Lin a, Wensheng Gan a , Philippe Fournier-Viger b , Tzung-Pei Hong c,d, Vincent S. Tseng Fast algorithms for mining high-utility itemsets with various discount strategies

[6] Souleymane Zida, Philippe Fournier-Viger, Jerry Chun-Wei Lin,Cheng-Wei Wu, Vincent S. Tseng, EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining