

Aplicación de Redes bayesianas usando Weka.

Cynthia Lorena Corso¹, Fabian Gibellini¹

¹Universidad Tecnológica Nacional, Facultad Regional Córdoba
Laboratorio de Sistemas de Información
Maestro M. López esq. Cruz Roja Argentina. Ciudad Universitaria S/N
cynthia@bbs.frc.utn.edu.ar
speakers@bbs.frc.utn.edu.ar

Abstract. Este trabajo tiene el objetivo de exponer la aplicación de una técnica de minería de datos como las Redes Bayesianas, aplicadas a la resolución de una problemática relacionada del campo de la ingeniería como lo es el mantenimiento correctivo. En él se expone cual es la problemática y el porqué de la elección de esta técnica para la clasificación de ocurrencias y cuales son los resultados obtenidos de aplicar esta técnica de minería de datos usando software Weka.

Keywords: Minería de Datos, Técnicas de clasificación, Redes Bayesianas, Aprendizaje automático, Weka.

1 Introducción

La minería de datos es un proceso que tiene como objetivo descubrir y extraer información relevante de base de datos o de otras fuentes de almacenamiento de datos, facilitando la identificación de patrones, tendencias como así también develar hechos anormales que pueden estar sucediendo.

Esto permite el aprovechamiento del valor de la información para que los directivos tengan un mejor conocimiento de su negocio y poder tomar decisiones más confiables.

Este trabajo tiene como objetivo la aplicación de una técnica de minería de datos dentro del grupo de técnicas de clasificación como lo es las redes bayesianas que será aplicada a una problemática perteneciente al ámbito de la industria, en una problemática muy frecuente como lo es el tema del mantenimiento correctivo de los maquinarias involucradas con el proceso productivo.

La problemática expresada por la empresa es la preocupación por la frecuencia en que se reportan órdenes de mantenimiento correctivo de las áreas involucradas en el proceso productivo.

Para develar cual pueden ser las causas de este comportamiento, como ya se ha expuesto aplicaremos la técnica de redes bayesianas y el clasificador seleccionado es NaivesBayes usando el software Weka.

El eje del estudio y análisis se centra en determinar cuáles son los posibles factores que tienen una mayor incidencia en los reportes de órdenes de mantenimiento correctivos generados en la empresa.

1.1 Redes Bayesianas: Fundamentación teórica.

El avance importante que ha tenido el campo de la tecnología y el abaratamiento de costos ha traído como consecuencia un aumento significativo en la cantidad de datos que son almacenados en muchas ocasiones en diferentes formatos.

La minería de datos es un mecanismo que nos permite facilitar la búsqueda de información valiosa en grandes volúmenes de datos. Este trabajo tiene como objetivo aplicar una técnica predictiva al campo de la industria como lo es el mantenimiento.

Las redes bayesianas es una técnica que pertenece al grupo de las técnicas de clasificación y consiste en un modelo gráfico que utiliza arcos para formar una gráfica acíclica y es aplicado en aquellas situaciones en que la incertidumbre se asocia con un resultado que se puede expresar en términos de probabilidad. Es decir los nodos del grafo representan variables, los arcos representan dependencia condicional y una distribución de probabilidad.

[1] En un comienzo, estos modelos eran construidos a mano basados en un conocimiento experto, pero en la actualidad se han investigado técnicas para aprender de dichos datos, tanto la estructura como los parámetros asociados al modelo.

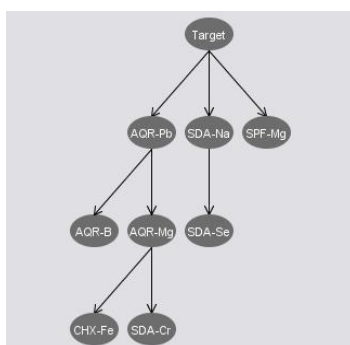


Fig. 1. Ejemplo de red bayesiana.

Esta técnica busca determinar relaciones causales que expliquen un fenómeno y es aplicado en aquellos casos que son de carácter predictivo.

[2] Es decir el razonamiento probabilístico o propagación de probabilidades consiste en difundir los efectos de la evidencia por medio de la red para conocer la probabilidad a posteriori de las variables. Es decir a determinadas variables (conocidas) se les otorga una probabilidad y en base a esto se obtiene una probabilidad posterior.

Hay ciertos conceptos básicos que están asociados a esta técnica como:

Grafo: Par de conjuntos $G=(X, L)$ donde X es un conjunto finito de elementos (nodos) y L es un conjunto de arcos.

Arco: subconjunto de pares ordenados.

Grafo Dirigido: Par ordenado $G=(X, L)$ donde X es el conjunto de nodos y L conjunto de arcos.

Grafo Acíclico: grafo que no tiene ciclos.

Una red bayesiana G define una distribución de probabilidad conjunta única sobre U dada por:

$$PB(X_1, X_2, \dots, X_n) = \prod PB(X_i | \pi X_i) \quad (1)$$

[3] Cualquier sistema de clasificación de patrones se basa en lo siguiente: dado un conjunto de datos (que dividiremos en dos conjuntos de entrenamiento y de test) representados por pares <atributo, valor>, el problema consiste en encontrar una función f(x) (llamada hipótesis) que clasifique dichos ejemplos.

La idea de usar el teorema de Bayes en cualquier problema de aprendizaje automático es que podemos estimar las probabilidades a posteriori de cualquier hipótesis consistente con el conjunto de datos de entrenamiento para así seleccionar la hipótesis más probable.

Para estimar estas probabilidades se han propuesto numerosos clasificadores bayesianos.

[2] Un clasificador en general suministra una función que clasifica una instancia especificada por una serie de características o atributos, en una o en diferentes clases predefinidas.

En general los clasificadores bayesianos son ampliamente utilizados debido a que presentan ciertas características:

- Son simples de construir y comprender.
- El proceso de inducción a partir de los mismos es veloz y sencillo.
- Robusto en cuanto considera atributos irrelevantes.
- Considera una importante cantidad de atributos para generar la predicción final.

Un clasificador bayesiano puede ser tomado en cuenta como un caso particular de una red bayesiana, en la que hay una variable que cumple el rol de la clase y los demás variables son consideradas atributos. La estructura va a depender fundamentalmente del tipo de clasificador.

Los clasificadores de redes bayesianas son:

Clasificador Bayesiano Simple (Naives Bayes classifier, NBC): permite obtener la probabilidad posterior de cada clase C_i , usando la regla de Bayes.

Este clasificador asume que los atributos son independientes entre sí dada la clase, así que la probabilidad se puede obtener por el producto de las probabilidades condicionales individuales de cada atributo.

La representación gráfica puede darse como una red bayesiana en forma de estrella. Es decir un nodo raíz, que representa la variable de la clase y que está conectada a los atributos.

Este clasificador bayesiano tiene extensiones y el fundamento de su uso es cuando se disponen de atributos que son dependientes.

Una manera de considerar esta dependencia es extendiendo la estructura básica de NBC, incorporando arcos a dichos nodos.

Las posibilidades básicas son:

- TAN: clasificador bayesiano simple aumentado con un árbol.
- BAN: clasificador bayesiano simple aumentado con una red.

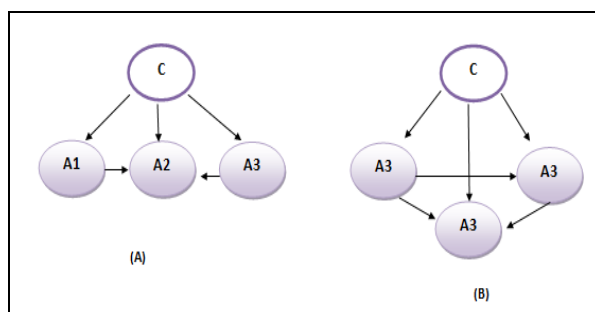


Fig. 2. Extensiones del clasificador bayesiano simple: (A) TAN, (B) BAN.

Redes bayesianas dinámicas: este clasificador permite representar el estado de las variables en un cierto momento de tiempo. En el caso de existir la necesidad de representar estos procesos dinámicos existe esta extensión conocida como red bayesiana dinámica (RBD).

Para las redes bayesianas dinámicas, generalmente se hacen las siguientes suposiciones:

- Proceso markoviano: el estado actual solo depende del estado anterior.
- Proceso estacionario en el tiempo: las probabilidades condicionales en el modelo no se alteran con el tiempo.

El aprendizaje de las redes bayesianas consiste en inferir un modelo, estructuras y parámetros, a partir de los datos, que puede ser agrupado en dos aspectos:

Aprendizaje Estructural: obtiene la estructura de la red bayesiana (o topología de red) tomando como punto de partida una base de datos, es decir las relaciones de dependencia entre las variables involucradas.

De acuerdo al tipo de estructura, podemos dividir los métodos de aprendizaje estructural en: Aprendizaje de árboles, Aprendizaje poliarboles, Aprendizaje de redes multiconectadas.

Aprendizaje Paramétrico: dada una estructura, obtiene las probabilidades asociadas. El requisito fundamental para llevar a cabo la tarea de aprendizaje de redes bayesianas es disponer de una base de datos en la que este detallado el valor de cada variable en cada uno de los casos.

En el caso de nuestra situación problemática considerada la elección de esta técnica de minería de datos se fundamenta en que en base a una red ya construida, y dados los valores concretos de algunas variables de una instancia, podrían tratar de estimar se los valores de otras variables de la misma instancia aplicando razonamiento probabilístico.

1.2 Metodología

La propuesta metodológica de trabajo consiste en el uso de una herramienta para el aprendizaje automático denominada Weka.

Esta herramienta es de distribución libre, desarrollada en Java y permite la implementación de técnicas de clasificación, agrupamiento y asociación; como así también realizar tareas de preprocesado y filtrado de datos.

A continuación se detallan las etapas llevadas a cabo para el análisis y estudio de la problemática expuesta anteriormente:

○ Selección de las variables significativas: En todos los estudios que se aplican técnicas de minería de datos, no siempre todas las características o atributos disponibles son realmente relevantes para obtener un modelo de conocimiento. Nuestro caso no es la excepción, el problema se enfoca en saber cuál es la probabilidad de que exista reportes de mantenimiento en una determinada maquinaria dadas ciertas condiciones (variables).

Es por ello que nos ayudaremos del software Weka para no tener en cuenta ciertas variables usando los filtros establecidos en la sección de Preprocesamiento.

A continuación se detallan cuales han sido las variables consideradas para nuestro estudio:

Tabla 1. Variables consideradas para la aplicación del algoritmo BayesNet.

Atributo	Descripción
fallo	Este atributo corresponde a la clase, que permitirá conocer que atributos inciden en la probabilidad de ocurrir un fallo en una maquinaria. El tipo de datos es booleano.
Fallo humano	Representa si el fallo de una maquinaria es responsabilidad humana o no. El tipo de datos es booleano.
Turno	Indica en que turno de trabajo se ha reportado el fallo. Este tipo de dato es nominal.
Mantenimiento Programado	Significa que en el momento que se reporta el fallo, existe la planificación de un mantenimiento programado. El tipo de dato es booleano.
Área	Indica el área de la empresa en la que se reporta el fallo. Este tipo de dato es nominal.

Es importante mencionar que la información histórica reportada por la empresa relacionada con la registración de órdenes de mantenimiento correctivo, se encuentra almacenada en formato de planilla de cálculo (Excel).

Además para reducir el conjunto de atributos seleccionados hemos usado un Filtro de Weka que se encuentra en la categoría de Atributos denominado “Remove”, que permite borrar un conjunto de atributos especificados en el archivo de datos.

- Transformación de los datos en el formato adecuado: En nuestro caso la fuente de datos obtenida se encuentra almacenada en una planilla de cálculo; el software que utilizamos es Weka por lo tanto es necesario que los datos tengan el formato adecuado para ser procesados por la herramienta. El formato adecuado de archivo .arff y la estructura del mismo se ilustra con el siguiente ejemplo:

(A)@relation mantenimiento

(B)

```
@attribute fallo {yes,no}
@attribute turno {manana,tarde,noche}
@attribute falloHumano {TRUE,FALSE}
@attribute mantenimientoProg {TRUE,FALSE}
@attribute area {Produccion,Embalaje}
```

(C)

```
@data
yes,tarde,TRUE,TRUE,Produccion
no,manana,FALSE,TRUE,Embalaje
yes,noche,FALSE,TRUE,Produccion
no,manana,FALSE,FALSE,Produccion
yes,noche,TRUE,TRUE,Produccion
no,manana,FALSE,TRUE,Embalaje
```

(A): En esta sección se define el nombre de la relación.

(B): Se especifican los atributos de la relación y el tipo de dato.

(C): Es la sección de datos propiamente dicha.

- Tratamiento de los valores faltantes en algunos atributos: Es importante una vez importado los datos a la herramienta realizar un análisis de cuál es el nivel de valores faltantes en las variables consideradas, ya que es un factor importante que puede llegar a influir en el modelo de predicción a obtener.

Weka nos ofrece una herramienta para el tratamiento de valores faltantes, que consiste en eliminar todas las instancias con valores nulos que es la que hemos considerado.

Weka permite aplicar una gran diversidad de filtros sobre los datos, permitiendo realizar transformaciones sobre ellos de todo tipo. En este caso hemos seleccionado un filtro que se encuentra en la categoría de Atributos cuya denominación es “ReplaceMissingValues”, que permite reemplazar todos los valores indefinidos por la moda en el caso de que sea el atributo nominal, como nuestro caso.

- Selección del algoritmo a utilizar: Para poder usar algún algoritmo perteneciente a la familia de los clasificadores bayesianos, se encuentra en la ruta `weka.classifiers.bayes.NaiveBayes`.

En este estudio se ha decidido elegir el algoritmo BayesNet. En la figura 3 se puede visualizar cuales son los parámetros que se pueden configurar para la aplicación de la técnica.

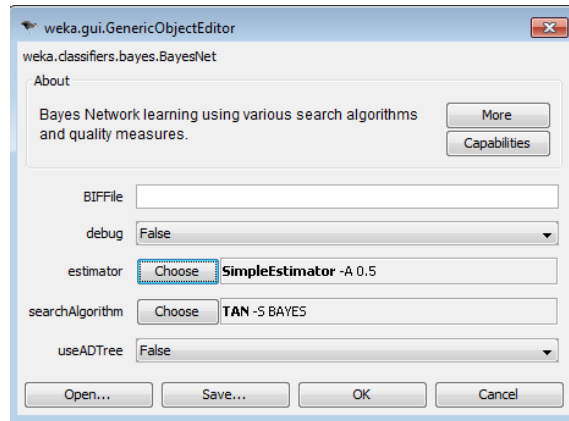


Fig. 3. Parámetros a configurar para la aplicación del algoritmo.

Antes de ejecutar el algoritmo se ha seleccionado una de las opciones de test: Cross-validation: esto permite realizar la evaluación mediante la técnica de validación cruzada; permitiendo establecer el número de muestras a utilizar.

Luego especificamos del conjunto de atributos seleccionados el que será considerado como clase principal, que será Fallo {yes/no}. Una vez que se ejecuta el algoritmo se visualiza en la ventana de salida con la siguiente información:

```

=== Summary ===
Correctly Classified Instances      30           75 %
Incorrectly Classified Instances    10           25 %
Kappa statistic                    0.5
Mean absolute error                 0.3337
Root mean squared error             0.4287
Relative absolute error             66.7452 %
Root relative squared error         85.7355 %
Total Number of Instances          40

=== Detailed Accuracy By Class ===
                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.7      0.2      0.778     0.7     0.737     0.766    yes
                0.8      0.3      0.727     0.8     0.762     0.766    no
Weighted Avg.   0.75     0.25     0.753     0.75    0.749     0.766

=== Confusion Matrix ===
 a  b  <-- classified as
14  6  | a = yes
 4 16 | b = no

```

Fig. 4. Ventana del clasificador de salida.

Weka nos brinda la posibilidad de visualizar el gráfico del árbol generado a partir de la clasificación.

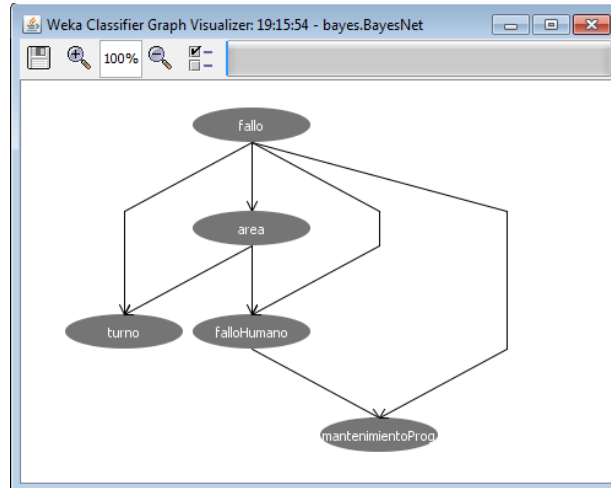


Fig. 5. Árbol de clasificación generado por Weka.

- Interpretación del modelo: consiste en analizar cuál es la probabilidad de las variables consideradas, respecto a la variable considerada como clase. Este aspecto se detallará más en profundidad en la sección de Resultados obtenidos y esperados.
- Validación del modelo de predicción: En esta etapa se consideran ciertos parámetros (que se detallan más adelante), para validar cual es el nivel de confianza del modelo de clasificación obtenido.

1.4 Resultados y Conclusiones

Una vez obtenido el árbol de clasificación, es necesario poder evaluar la calidad o nivel de confianza del mismo. La métrica Kappa Statistic es considerada para tal fin. El Kappa Statistic es una medida que permite medir el nivel de predicción respecto a la variable considerada como clase.

En los resultados obtenidos en este estudio se da para cada variable (cada nodo del árbol) el nivel de probabilidad de la misma respecto a la clase principal. Por ejemplo: Si seleccionamos el atributo área en el árbol, nos muestra información que se resume a continuación:

fallo	Produccion	Embalaje
yes	0,69	0,31
no	0,31	0,69

Fig. 6. Tabla de distribución de probabilidad para la variable área.

De acuerdo a los resultados que figuran en la tabla para la variable **área** podemos inferir que si el área es Producción la probabilidad que se produzca un reporte de fallo en una maquinaria es del 69% y en el caso de tratarse del área de Embalaje será del 31%.

Realizando este procedimiento con todas las variables consideradas se concluye que:

- Área de producción en el turno de la tarde hay una probabilidad del 74% que se reporte un fallo en una maquinaria del sector.
- En el área de producción existe una probabilidad del 70% que los reportes de fallos a maquinaria se deben por falla humana o error en la operación.

Hay una clara tendencia de que la mayor frecuencia de reporte de mantenimiento se origina en el área de Producción básicamente en el turno tarde, con lo que se agrega un porcentaje alto de los reportes se deben a fallos humanos o error en la operación. Por lo cual se sugieren las siguientes recomendaciones:

- Capacitación relacionada con el manejo y funcionamiento de las maquinarias, orientada al sector de Producción específicamente a los operarios del turno tarde.
- Analizar las condiciones ambientales y físicas del sector de producción como ubicación, luminosidad y posibles ruidos que puedan interferir en el trabajo.

Estos factores, que no están relacionados directamente con el trabajo, pero se ha comprobado que afectan a la productividad con la probabilidad de generar malas operaciones en las maquinarias.

- Analizar las condiciones ambientales actuales de las maquinarias del área de producción como luminosidad, nivel de humedad, instalaciones eléctricas sean las adecuadas. Es posible que esto puede estar afectando mal funcionamiento en las misma generando reportes de mantenimiento.

En este trabajo se ha planteado una propuesta metodológica de un caso práctico de minería de datos usando redes bayesianas. Una vez generado el modelo se han propuesto una serie de recomendaciones con el objetivo de identificar las variables que tienen un mayor nivel de incidencia en la problemática expuesta.

Los métodos bayesianos realizan un aporte desde el punto de vista cuantitativo, es decir da una medida probabilística de las variables consideradas en una determinada situación. Esta es una de las diferencias fundamentales del uso de las redes bayesianas respecto a otros métodos como redes neuronales y los arboles de decisión, que no ofrecen una medida cuantitativa de la clasificación.

Referencias

1. Basilio Serra; Araujo.: Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software Weka; Prentice Hall (2006)
2. Luque Malagón; Constantino.: Clasificadores Bayesianos. El algoritmo de Naives Bayes. (2003)
3. Villanueva Velasco; David.: Redes Bayesianas. Inteligencia Artificial II. (2007)
4. Weka 3. Data Mining Software in Weka, <http://www.cs.waikato.ac.nz/ml/weka/>
5. Portugal, Roberto; Carrasco, Miguel.: Ensamble de algoritmos bayesianos con árboles de decisión; Departamento de Ciencia de Computación; Universidad Católica de Chile.
6. García; Morate.: Weka Tutorial, <http://www.metaemotion.com/diego.garcia.morate>

7. Porta Zamorano; Jordi.: Técnicas cuantitativas para la extracción en un corpus; Escuela Politécnica Superior, Universidad Autónoma de Madrid (2006)
8. Ordieres Meré; Juan, Martínez de Pisón Ascacibar; Fco. Javier.: Data mining in industrial processes; Area de proyectos de Ingeniería, Universidad de la Rioja/Universidad de León.
9. Bregón; Aníbal, Arancha; Simón, Alonso; Carlos, Belarmino; Pulido, Moro; Isaac.: Un sistema de razonamiento basado en casos para la aplicación de fallos en sistemas dinámicos, Departamento de Informática; Universidad de Valladolid.
10. Cruz; Antonio M., Barr; Camerón, Castilla Casado; Norberto.: Evaluación de las solicitudes de mantenimiento correctivo usando técnicas de agrupamiento y clasificación, Revista de Ingeniería Biomédica, volumen 2, Pág 65-76 (2008)