Métodos Bayesianos II

García Prado, Sergio
sergio@garciparedes.me

3 de mayo de 2017

Resumen

En este documento se realiza una descripción acerca de los algoritmos de clasificación basados en generación de Redes Bayesianas *y el* Teorema de Bayes *referido a la probabilidad condicionada. Además, se realizan varios experimentos utilizando las estrategias de construcción de estructura de redes bayesianas descritas previamente, sobre conjuntos de datos de distintos tamaños mediante tests de* Validación Cruzada *y un conjunto de datos formado por instancias independientes del resto.*

1. Redes Bayesianas

"Una red bayesiana es un modelo gráfico probabilístico que representa un conjunto de variables aleatorias y sus dependencias condicionales a través de un grafo acíclico dirigido. Por ejemplo, una red bayesiana puede representar las relaciones probabilísticas entre enfermedades y síntomas. Dados los síntomas, la red puede ser usada para computar la probabilidad de la presencia de varias enfermedades." [Wik17]

"Formalmente, las redes bayesianas son grafos dirigidos acíclicos cuyos nodos representan variables aleatorias en el sentido de Bayes: las mismas pueden ser cantidades observables, variables latentes, parámetros desconocidos o hipótesis. Las aristas representan dependencias condicionales; los nodos que no se encuentran conectados representan variables las cuales son condicionalmente independientes de las otras. Cada nodo tiene asociado una función de probabilidad que toma como entrada un conjunto particular de valores de las variables padres del nodo y devuelve la probabilidad de la variable representada por el nodo." [Wik17]

Existen algoritmos eficientes que llevan a cabo la inferencia y el aprendizaje en redes bayesianas. En este documento se habla de 3 de los algoritmos de generación de estructuras de red: *Naive Bayes* en la sección 1.1, *K2* en la sección 1.2 y *TAN* en la sección 1.3.

1.1. Estructura Naive Bayes

El algoritmo *Naive Bayes* genera la estructura de red suponiendo que la clase depende de todos los nodos de manera simple, por lo que se crea un árbol de altura 1 donde el padre es la variable aleatoria que representa la clase mientras que los hijos están formados por los nodos que representan las variables aleatorias de los atributos del conjunto de datos.

1.2. Estructura K2

El algoritmo de generación de redes *K2* se basa en la suposición de que existe un orden total entre el conjunto de atributos y la clase del conjunto de datos. Dicha ordenación se realiza a partir del estimador de máxima similitud con corrección bayesiana, que se diferencia del de Laplace en que este varía para cada nodo. La estructura de la red se crea suponiendo que para el nodo X_k, puede tener como padres al subconjunto de nodos $X_1, ..., X_{k-1}$ en el caso de que este mejore su puntuación.

1.3. Estructura TAN

El algoritmo *TAN* (o *Tree Aumented NaiveBayes*) genera la estructura de red en dos fases. La primera de ellas se corresponde con el algoritmo *Naive Bayes* descrito anteriormente en la sección 1.1. La segunda fase compone las dependencias entre atributos (no utiliza el nodo de la clase) mediante el algoritmo de *Chow y Liu*, que genera el *árbol recubridor máximo* a partir de la ecuación (1) (información mútua entre dos variables aleatorias condicionadas a la clase).

$$I_{\hat{Pr}}(X,Y|C) = \sum_{x,y} \hat{Pr}(x,y|C) log(\frac{\hat{Pr}(x,y|C)}{\hat{Pr}(x|C)\hat{Pr}(y|C)}) \quad (1)$$

2. Experimentos

Tras haber descrito las *Redes Bayesianas* en la sección anterior, a continuación se presentan los resultados obtenidos tras realizar un conjunto de experimentos sobre el conjunto de datos **Credit**, que se describen en la sección **??**. Los tests han consistido en la evaluación del comportamiento de los algoritmos de generación de *Redes Bayesianas* implementadas en *Weka* [too] variando el tamaño de los conjuntos de entrenamiento.

La metodología seguida ha sido la siguiente: Para cada conjunto de datos de entrenamiento se ha realizado un experimento de *Validación Cruzada de 10 particiones* a partir del cual se ha almacenado la tasa de error obtenida. Además, se han guardado los modelos generados en cada caso para después probarlos sobre un conjunto de datos independiente del de entrenamiento. Esta tarea se ha repetido con tres los conjuntos de datos de 3 tamaños diferentes sobre los algoritmos de generación de *Redes Bayesianas* descritos en la sección anterior: *Naive Bayes*, *K2* y *TAN*.

A continuación se describe la naturaleza del conjunto de datos utilizado para los experimentos así como los tamaños de sus particiones.

2.1. Conjunto de Datos

El conjunto de datos utilizado se denomina *Credit* y se puede acceder a él a través de `https://github.com/garciparedes/machine-learning-bayesian-2/tree/master/weka/datasets` [GP17]. Está formado por **11 atributos** de carácter nominal más la clase de destino formada por **2 valores**. Dichos atributos presentan un rango de valores reducidos, encontrandose todos ellos entre **2 y 4 valores distintos**.

En cuanto a los conjuntos de datos utilizados para los experimentos, todos ellos contienen los mismos atributos así como instancias representativas para todos los valores posibles de los atributos. Tampoco se dan atributos desconocidos. a continuación se describen los tamaños de cada conjunto de datos:

- **Datos_Credit_100**: Está formado por 100 instancias.
- **Datos_Credit_1000**: Está formado por 1000 instancias.
- **Datos_Credit_10000**: Está formado por 10000 instancias.
- **Test_Credit_1000**: Está formado por 1000 instancias.

Por último es necesario describir tanto el estimador como los parámetros de configuración utilizados para cada algoritmo. En cuanto al *método de estimación* de probabilidades, se ha escogido el *Estimador de Máxima Verosimilitud con corrección de Laplace* y $m = 0{,}5$. Dicho estimador se muestra en la ecuación (2), donde n_c representa el número de ejemplos de entrenamiento con la clase b_j, n el número de ejemplos de la de entrenamiento con la clase b_j y el atributo a_i. El valor p se obtiene mediante $p = Pr(A = a_i|B = b_j)$, es decir, la estimación a priori y m un determinado peso para la estimación a priori.

$$Pr'(A = a_i|B = b_j) = \frac{n_c + mp}{n + m} \quad (2)$$

En cuanto a los algoritmos de clasificación, *Naive Bayes* no tiene más parámetros adicionales. *K2* se ha configurado para que no comience de una red *Naive Bayes* y se ha permitido que cada nodo tenga un número arbitrario de padres. En el caso de *TAN*, se ha seleccionado el método de puntuación de *Máxima Verosimilitud con corrección de Laplace* descrito en la ecuación (2) con parámetro $m = 0{,}5$ al igual que en el caso del método de estimación de probabilidades.

En la sección 2.2 se presentan los resultados obtenidos así como una discusión acerca de los mismos.

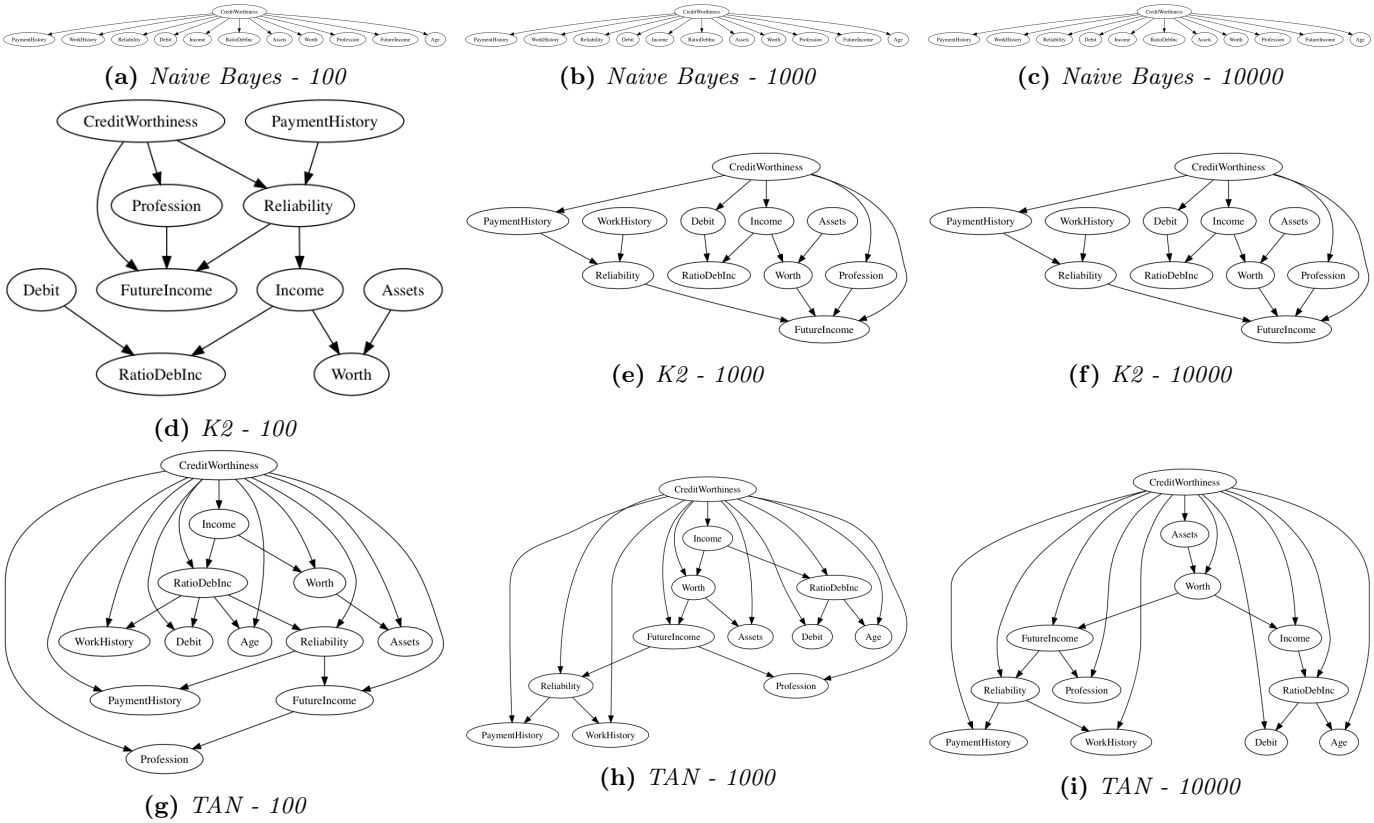
2.2. Resultados

En la figura 1 se muestra la estructura de las distintas *Redes Bayesianas* generadas por cada algoritmos dependiendo del tamaño del conjunto de datos de entrenamiento. Nótese que estas redes han sido generadas a partir de un experimento de *Validación Cruzada de 10 particiones* sobre conjuntos de datos de *100*, *1000* y *10000* instancias respectivamente.

En cuanto al algoritmo *Naive Bayes*, la estructura de la red generada es trivial en ambos casos, suponiendo dependencia directa de la clase sobre todos los atributos, por lo que se genera un árbol de profundidad 1 donde el nodo raiz es la clase y los nodos hoja se corresponden con los 11 atributos. Nótese que el tamaño del conjunto de datos de entrenamiento no genera variaciones tal y como se muestra en las figuras 1a, 1b y 1c.

Las redes generadas por el algoritmo *K2* presentan una característica diferenciadora respecto del resto. En este caso no todos los atributos tienen como padre al nodo referido a la *clase* (algunos a pesar de no ser la clase de destino no tienen padre). Esto se traduce de manera práctica en que dichos atributos no serán utilizados durante las fases de clasificación de nuevas instancias. La razón es la estrategia de construcción de redes seguida por este algoritmo, que añade una arista tan solo cuando la puntuación obtenida mediante el indicador de verosimilitud se mejora. Las estructuras generadas a partir de cada conjunto de datos de distinto tamaño se muestran en las figuras 1d, 1e y 1f.

El algoritmo *TAN* tal y como se ha descrito anteriormente, es una combinación del algoritmo *Naive Bayes* y el algoritmo *Chow Liu* obviando la clase de destino. Por tanto, debido a la primera fase de *Naive Bayes*, la clase depende de todos los atributos directamente. En la fase de *Chow Liu* se forman las relaciones entre atributos. Nótese que las estructuras de red formadas por *TAN* tienen muchas más aristas que el resto, lo cual añade complejidad en el cálculo de probabilidades para la clasificación de nuevas instancias. Las redes generadas se muestran en las figuras 1g, 1h y 1i.



(a) *Naive Bayes - 100* **(b)** *Naive Bayes - 1000* **(c)** *Naive Bayes - 10000*

(d) *K2 - 100* **(e)** *K2 - 1000* **(f)** *K2 - 10000*

(g) *TAN - 100* **(h)** *TAN - 1000* **(i)** *TAN - 10000*

Figura 1: *Redes Bayesianas generadas a partir de DatosCredit*

Clasificación mediante Redes Bayesianas									
Datos	Tasa de Error								
	Naive Bayes			K2			TAN		
	100	*1000*	*10000*	*100*	*1000*	*10000*	*100*	*1000*	*10000*
Entrenamiento	36,00 %	31,70 %	29,69 %	28,00 %	36,10 %	28,50 %	35,00 %	29,50 %	28,27 %
Datos Test	30,00 %	27,80 %	27,50 %	33,30 %	30,40 %	25,40 %	33,30 %	24,90 %	25,00 %

Tabla 1: *Tasas de error obtenida a partir de distintas configuraciones a nivel de estructura de* Redes Bayesianas

A continuación se describen los resultados a nivel de tasas de error obtenidas en los experimentos. Estos han sido resumidos en la tabla 1. Tal y como se dijo anteriormente, los resultados obtenidos en la fase de entramiento se corresponden con una metodología de *Validación Cruzada de 10 particiones*, mientras que los obtenidos en la fase de *Test* son relativos a la clasificación de las instancias del conjunto de datos de 1000 instancias a partir del modelo obtenido en la fase anterior.

La tendencia habitual de los tres algoritmos de clasificación es de reducir su tasa de error conforme el tamaño del conjunto de datos de entrenamiento aumenta. Sin embargo, se pueden apreciar dos casos atípicos (entrenamiento con *K2*). Estos resultados pueden ser debidos a distintos factores como la generación puntual de particionamientos que hagan caer en el peor caso posible.

A nivel de resultados, los tres algoritmos ofrecen tasas de error similares. Debido a las pequeñas variaciones a nivel de tasas de error entre ellos, no se puede asumir que ninguno ofrezca mejores resultados que el resto sobre los conjunto de datos utilizados.

Referencias

[CCAG17] Teodoro Calonge Cano and Carlos Javier Alonso GonzáLez. Técnicas de Aprendizaje Autómatico, 2016/17.

[GP17] Sergio García Prado. Métodos bayesianos 2. `https://github.com/garciparedes/machine-learning-bayesian-2`, 2017.

[too] Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.

[Wik17] Wikipedia. Bayesian network — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Bayesian%20network&oldid=776991271`, 2017. [Online; accessed 02-May-2017].