INVESTIGACIÓN OPERATIVA: Trabajo Final

Integrantes:
Juarez, Tomás
Fernández Cocirio, Mariano

Director:
Mg. Bueno, Moisés Evaristo

2016

Clasificador bayesiano ingenuo aplicado a reputación en e-commerce

Tomás Juarez
tomasjuarez.exa@gmail.com
Mariano Fernández Cocirio
marianococirio@gmail.com
Facultad de Ciencias Exactas-Universidad Nacional del Centro de la Provincia de Buenos Aires

1. RESÚMEN

En los tiempos modernos el comercio electrónico ha ganado un importante lugar entre los principales métodos para realizar operaciones de compra-venta. Uno de los inconvenientes que acarrean estos tipos de sistemas es la confiabilidad de los usuarios a la hora de realizar una compra. Este problema es mayor si nos situamos en el contexto de una ONG, cuyo objetivo es el de brindar una plataforma para el intercambio de bienes de forma gratuita. Por este motivo se propone una solución utilizando clasificadores bayesianos para optimizar la categorización de los usuarios, etiquetándolos de acuerdo a las valorizaciones obtenidas a partir de su historial en el sistema electrónico.

2. PALABRAS CLAVES

Comercio Electrónico, Red Bayesiana, Investigación Operativa.

3. INTRODUCCIÓN

Durante el gran avance de la tecnología en los últimos años, uno de los sectores que recibió el mayor impacto fue el comercio, el cual comenzó a transitar un nuevo paradigma, implicando la introducción del mercado en medios informáticos, lo que hoy conocemos como el comercio electrónico.

El comercio electrónico consiste en la compra/venta de productos o servicios, a través de una plataforma en la red. Esto da como resultado el que a través de internet dos usuarios desconocidos realicen transacciones libremente, con el problema de la confiabilidad, puesto que el comprador y vendedor no se conocen. En la mayoría de los sitios de comercio electrónico el usuario que produce un mayor riesgo es el vendedor, ya que usualmente el mismo envía su producto luego de haberse efectuado el pago.

En nuestro marco particular, tratamos con una ONG que se encarga de conectar otras organizaciones que ofrecen o necesitan bienes y servicios.

En este trabajo, nos hemos enfocado en etiquetar a un usuario vendedor según su historial en el sitio, técnica previamente implementada por otras plataformas de comercio electrónico de manera exitosa. Se ha desarrollado

entonces un clasificador de usuarios como optimización al proceso de compras. El mismo tiene como fin determinar la posición de un usuario específico en una categoría del espectro propuesto, de acuerdo a su accionar previo. El clasificador provee una escala de cinco categorías: muy bueno, bueno, regular, malo, muy malo, ordenada en orden decreciente según la confiabilidad.

El mismo se encuentra desarrollado en forma de clasificador bayesiano ingenuo [1]. Este se basa en la inferencia a una hipótesis a través del teorema de Bayes, suponiendo independencia entre las variables observadas. Es muy importante que las variables sean independientes, ya que esto permite acotar la cantidad de datos necesarios para realizar una inferencia exitosa, además de permitir utilizar ciertas técnicas que brindan mayor performance.

El objetivo de este trabajo es encontrar un método confiable de categorización de usuarios en un sitio de e-commerce, logrando etiquetarlos de forma acertada para prevenir posibles fraudes, y detectar usuarios que sólo generan problemas con sus interacciones en el sitio.

Primero se desarrollará un modelo de clasificador bayesiano ingenuo capaz de afrontar la problemática. Luego se estudiarán las distintas formas de generar resultados a partir de dicho clasificador, y sus formas de entrenamiento, para lograr discernir la más apropiada para la situación. Finalmente se implementará el mismo teniendo en cuenta la futura integración con la página de comercio electrónico en cuestión, brindando una interfaz sencilla, y su núcleo lógico en el lenguaje de programación web PHP.

4. DESARROLLO

4.1 Clasificador Bayesiano Ingenuo

4.1.1 Arquitectura

El clasificador bayesiano ingenuo se basa en una hipótesis que infiere, o es inferida por más de una variable o evidencia. Cada una de esas evidencias a su vez es independiente de las otras, además su valor puede ser provisto por observación de los hechos.

Una representación general del mismo se puede ver en la figura (1), en donde podemos observar una hipótesis actuando sobre diversas evidencias. Nótese que el sentido de la conexión delata el sentido de la inferencia, si se quisiera partir de la evidencia para determinar la probabilidad de una hipótesis la dirección de las conexiones debería ser la inversa.

A partir del modelo antes mencionado se puede comenzar a detallar los nodos, los mismos cuentan con una tabla de probabilidades condicionales a priori,

es decir, probabilidades detectadas a partir de la recolección de datos, que demuestran el comportamiento del nodo respecto a los valores de quienes lo infieren.

Los nodos hoja, es decir los que no infieren a ningún otro nodo, obtienen normalmente sus valores a partir de la observación. Este valor es conocido como likelihood, el cual es el valor de la probabilidad observada en el sujeto a analizar.

Cada nodo a su vez cuenta con diversos valores asignables por quién configure la red, en los mismos se encuentran todas las alternativas que puede tomar la evidencia, y una distribución probabilística asociada a los mismos. Dicha distribución debe contar con ciertas características, como que el área bajo su curva sea 1.

Luego, a partir de las distribuciones que determinan el valor a priori en los nodos, y el likelihood, se puede inferir utilizando el teorema de Bayes para calcular el valor a posteriori de un nodo, dando así la probabilidad de ocurrencia de cada valor para un nodo determinado. En nuestro caso lo que interesará será el valor que determine la categoría del usuario sobre el cual se tomen los datos observados. El resultado se dará de acuerdo a una decisión a partir de las probabilidades de ocurrencia de cada valor.

4.1.2 Entrenamiento

El entrenamiento puede ser llevado a cabo de dos formas, alimentando el clasificador con un conjunto de datos ya clasificados previamente, y a partir de estos generar una convergencia en los valores de las probabilidades condicionales, o bien tomando una tabla representativa de condicionales para cada nodo, la cual permite una clasificación efectiva en casos de test pre-diseñados.

Ambos métodos pueden ser acompañados con un entrenamiento dinámico a medida que el clasificador es utilizado, esto resulta mejor en el caso que las distintas entradas presentan una gran heterogeneidad.

La forma de conseguir la convergencia de una probabilidad es utilizando datos preexistentes en los cuales todos los valores se encuentren definidos, de esta forma los mismos se procesan consiguiendo las probabilidades de ocurrencia, cuando llega el momento en que al añadir un nuevo dato al clasificador no modifica las probabilidades significativamente, eso significa que se ha llegado a su punto de convergencia.

Una gran ventaja de estos tipos de clasificadores es la simplicidad de entrenamiento, para lograr esto es que se debe pedir que las distintas evidencias

sean independientes entre sí, ya que de no serlo se generarían condicionales complejas, y ciclos en el modelo, lo que elevaría la complejidad del mismo.

4.2 Metodología de trabajo

4.2.1 Flujo de trabajo

Se presenta el flujo de trabajo que se utilizó para desarrollar ordenadamente el proyecto. Los pasos que se siguieron fueron los siguientes:

1. Detección de variables.
2. Recolección de información.
3. Pre-procesamiento de la información.
4. Planteamiento del modelo para el clasificador.
5. Implementación del clasificador

Se procederá a explicar en detalle cómo fueron realizados los distintos pasos y cuáles fueron las herramientas y metodologías utilizadas.

4.2.2 Detección de variables

Una parte fundamental a la hora de realizar un clasificador bayesiano ingenuo es la detección de las variables, ya que las mismas determinarán los resultados, además también se debe certificar la independencia de las mismas.

En este caso las variables detectadas a partir de la información provista por la página, y que cumplían con la independencia previamente mencionada fueron tres: Votos, Tiempo en página, y Tiempo respuesta de mensajes.

La variable Votos simboliza el porcentaje de votos negativos y positivos sobre el total de votos que posee, los valores que puede tomar no son mutuamente excluyentes, ya que un usuario puede tener 20% de votos negativos, y 80% de votos positivos. Los votos son computados a través de las valoraciones previas que han dado usuarios que han recibido o prestado algún bien al usuario. Es considerada una variable de interés ya que si un usuario tiene una tendencia a ser fraudulento, puede llegar a serlo en un futuro, mientras que un usuario que mantiene una conducta intachable es menos probable que cometa un fraude.

La variable Tiempo en página representa el tiempo que el usuario ha estado registrado en la página, ya que se considera que un usuario con mayor tiempo de actividad en la página tiene una menor tendencia a realizar acciones que perjudiquen su reputación, esto se da en cierta forma para evitar usuarios que se registran, realizan unas pocas operaciones exitosas, y ya poseen el mismo nivel que un usuario que ha estado realizando una excelente labor durante años. El tiempo en la página está simbolizado por rangos de valores excluyentes, es decir, un usuario no puede haber estado entre 2 meses y 6 meses, y más de 8 meses al mismo

tiempo. Los valores posibles son: Menos de dos meses, Entre dos y ocho meses, y Más de ocho meses.

Por último la variable Tiempo respuesta de mensajes se utiliza para denotar el tiempo que tarda un usuario en responder consultas, esto se considera una variable de interés ya que es premiable el hecho de que un usuario se interese y responda rápidamente las consultas, agilizando las negociaciones. De la misma forma que la variable anterior, esta variable posee valores excluyentes. Los valores posibles son: Menos de un día, Entre uno y cinco días, y Más de cinco días.

Con estas variables se realizará la inferencia sobre la categoría del usuario, la cual es nuestra hipótesis, que puede tomar cinco valores: Muy Malo, Malo, Regular, Bueno, Muy Bueno. De los cuales será seleccionado el que tenga mayor probabilidad de ocurrencia luego de la inferencia a partir de los valores observados de las otras tres variables.

4.2.3 Recolección de información

La recolección de la información se realizó de forma manual, a través de distintas páginas reconocidas de comercio electrónico a nivel mundial. Esto se hizo al observar los datos y el rango asignado en cada una de estas páginas a los mismos.

Esta etapa consumió gran parte del trabajo, ya que el conjunto de datos necesarios para llegar a una convergencia con un umbral despreciable debe ser muy grande. Finalmente se decidió finalizar la recolección de datos al conseguir un conjunto de datos decente, con un umbral quizás no despreciable, pero sí aceptable a fines prácticos.

4.2.4 Pre-procesamiento de la información

Luego de haber recaudado la información es el momento de confeccionar las tablas de probabilidades condicionales asociadas a cada nodo. Esta transformación de los datos es necesaria para la manipulación de clasificadores bayesianos ingenuos, ya que los cálculos a efectuar se basan en las probabilidades de ocurrencia a priori, y las probabilidades observadas.

Para los nodos correspondientes a evidencia las tablas fueron dadas por la equiprobabilidad de ocurrencia, ya que los valores de las mismas serían obtenidos a partir de la observación de un usuario.

Mientras tanto la tabla con mayor dificultad a confeccionar es la referida a la hipótesis, es decir, al nodo que nos determina la categorización. Luego de analizar el conjunto de datos recolectado, y hallar una convergencia aceptable, la tabla de

probabilidades quedó determinada como la TABLA 1, con el valor en cada entrada en probabilidad de ocurrencia porcentual.

Una particularidad observable en este punto es el castigo más severo ante las malas acciones que el premio ante las buenas acciones, se puede ver esto al observar distintas entradas que muestran ante acciones positivas probabilidades de quedar categorizado como Regular, Bueno, o Muy Bueno, mientras que ante acciones negativas hay una mayor densidad sobre Malo, o Muy Malo.

4.2.5 Planteamiento del modelo para el clasificador

El modelo planteado se basa en la inferencia de las variables detectadas sobre la hipótesis, el mismo se puede apreciar en la figura (2), el mismo fue materializado para fines de prácticos y de evaluación en Netica [2].

Bajo este modelo se corrieron pruebas con datos cuyos resultados eran sabidos, para determinar si se estaba realizando una correcta clasificación, el resultado es mostrado en el nodo Calificación, donde el valor con mayor probabilidad de ocurrencia será el resultado de la red bayesiana ante un sujeto observado.

Una parte fundamental del planteo del modelo es determinar la dirección de la inferencia, ya que esto determina cuál será la entrada, y cuál la salida de nuestro clasificador.

4.2.6 Implementación del clasificador

El clasificador fue implementado en el lenguaje de programación PHP, de forma tal que pueda utilizarse en un contexto web. La implementación se realizó según los principios de la programación orientada a objetos, puesto que debe ser flexible respecto a la modificación así como también manteniendo la legibilidad del código fuente. La figura (8) muestra el diagrama de clases asociado.

Puesto que por el momento no poseemos acceso al servidor en donde se aloja el sitio web de la ONG, el clasificador debe poder ser operado mediante una interfaz sencilla de utilizar, logrando así una mayor facilidad a la hora de su integración

El clasificador tiene un método público, cuyos parámetros son *porcentaje de votos positivos (float), tiempo medio en responder mensajes de otros usuarios en días (float) y tiempo que llevan en el sitio (integer)*. Para representar la tabla de probabilidades multidimensional, se optó por insertar una pequeña tabla en cada instancia de *Label* en donde representa la columna de la etiqueta en cuestión (muy malo, malo, neutro, bueno y muy bueno) con todas las posibles entradas. Cada entrada es definida mediante una 3-tupla o triada, instancias de la clase *Triad*. A su

vez, cada triada tiene -valga la redundancia- tres variables. Nótese que las variables pueden establecerse por rangos, en donde hay una cota mínima y una cota superior (mayores o iguales a X y menores a Y) o simplemente una cota máxima (todos los menores a X). En el caso de las variables concretas, sólo tienen un valor comparable mediante una igualdad. Cada entrada de la tabla es, en efecto, una instancia de *Triad* como clave de un hashmap y un valor de punto flotante como valor del mismo representando la probabilidad asociada.

Dado que se utiliza el paradigma orientado a objetos, la versión de PHP debe ser mayor a 5.3.

4.2.7 Software

Pese a que el clasificador en sí mismo no cuenta con interfaz gráfica ya que es un componente listo para su integración en un sistema, se adicionó un ejemplo de uso, en el cual se requieren los resultados del clasificador para clasificar un usuario cargado por medio de una interfaz gráfica simple, en la cual posteriormente los resultados son visualizados[3].

La interfaz de usuario de esta aplicación se puede ver en la figura (9), en la misma se requieren los datos necesarios para el correcto funcionamiento del clasificador, los cuales son: Cantidad de votos positivos, cantidad de ventas totales, tiempo de respuesta, y tiempo en el sitio.

Luego que los datos son cargados correctamente se procede a mostrar los resultados de la clasificación, en la figura (10) se puede ver el resultado de un usuario con 500 ventas realizadas, 400 votos positivos, que responde en 0.5 días, y lleva 6 meses en el sitio, el resultado de la clasificación es Excelente, ya que es el resultado con mayor probabilidad.

Otra peculiaridad a notar es la desambiguación entre la escala del clasificador en su modelo, y cómo desea utilizarlo luego quién lo requiera, ya que en el modelo se plantea un espectro de cinco valores, con dos valores extremos, un neutro, y dos intermedios, que luego el usuario le agrega desde su aplicación la semántica que prefiera.

4.3 Resultados y análisis

Los resultados serán mostrados en base a la herramienta Netica, ya que permite una interpretación más sencilla de los mismos al contar con un entorno gráfico, mientras que el clasificador implementado sólo cuenta con una interfaz que a partir de la evidencia de entrada, retorna una clasificación como salida.

Primero podemos ver la clasificación de un usuario con un 80% de votos positivos, que responde en menos de un día las consultas, y lleva entre dos y ocho

meses en la página, figura (3), el resultado se interpreta como un usuario Muy Bueno, ya que dicho valor posee la mayor probabilidad.

También podemos ver un caso distinto, en alguien nuevo (menos de dos meses en la página), que tarda menos de un día en contestar las consultas, y posee un 70% de votos negativos, figura (4), quedando dentro de la categoría Malo, esto se da porque a pesar de los votos negativos realiza las respuestas en un tiempo muy corto. Pero si variamos el tiempo de respuesta por entre uno y cinco días, queda dentro de la categoría Muy Malo, figura (5).

Para ver un ejemplo cuya decisión es más reñida, podríamos observar un usuario cuyos votos están igualados entre negativos y positivos, pero además responde en menos de un día, y lleva entre dos y ocho meses en la página, FIGURA 6. Termina por entrar en Muy Bueno por una diferencia mínima, ya que responde velozmente, estos casos extremos son los que con una convergencia cuyo umbral fuera más pequeño tendrían menos probabilidades de clasificar incorrectamente, de todas formas se puede apreciar que la incertidumbre no incluye los valores negativos entre los que pelean por ser la decisión final.

También hay que recordar a la hora de analizar los resultados que la salida del clasificador es una decisión, sin importar el porcentaje de la variable discreta asociada, es decir, alguien con 1000 ventas realizadas exitosamente, es decir, 1000 votos positivos, que contesta en 0.1 horas, y lleva 5 años en nuestra página tendrá la salida mostrada en la figura (7). Mientras que alguien con 9 meses de antigüedad, que contesta en 20 horas, y tiene 2 ventas realizadas exitosamente, con 2 votos positivos, tendrá el mismo resultado, esto se da por la discretización utilizada para los datos, y el resultado es correcto, ambos son usuarios Muy Bueno, por lo que la decisión pasa a ser acertada.

Los resultados anteriormente analizados pueden ser probados también en la herramienta online, la única diferencia existente es que en la misma se ve el resultado obtenido, pero sin ver los nodos que causan la inferencia, lo cual es considerado apropiado en dicho entorno, pero dentro del ámbito del informe es mejor mostrar la red completa a la hora de realizar la decisión.

5. CONCLUSIONES

Luego de la realización del trabajo el grupo se encuentra conforme dado al éxito de implementación, y planteo del clasificador buscado. Se logró llegar a una solución al problema original de la confiabilidad de los usuarios en la ONG destinada a conectar con otras organizaciones que no se conocen por medio de un sitio de comercio electrónico, logrando dar un porcentaje de confiabilidad en base a las actuaciones del usuario en acciones previas. De esta forma ya no es relevante quién es el vendedor, si no cómo vende, es decir, la calidad como vendedor,

permitiendo al prestador de un servicio, o a quien lo requiere, poder discernir entre distintos usuarios al realizar la selección del más indicado.

Además el mismo se realizó utilizando técnicas básicas de teoría de la probabilidad, inteligencia artificial y minería de datos.

Como trabajo futuro queda abierto el agregar más variables al clasificador, generando una mayor exactitud, y también agregar la relación entre las distancias, los tiempos de entrega, y la reputación del usuario, para poder obtener la mejor elección a la hora de requerir, o brindar un servicio, o material.

6. TABLAS/FIGURAS

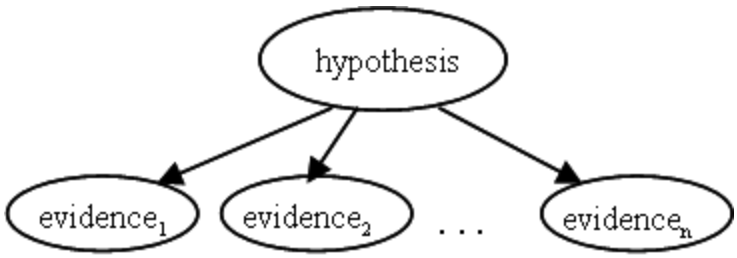

FIGURA 1 - Clasificador bayesiano ingenuo

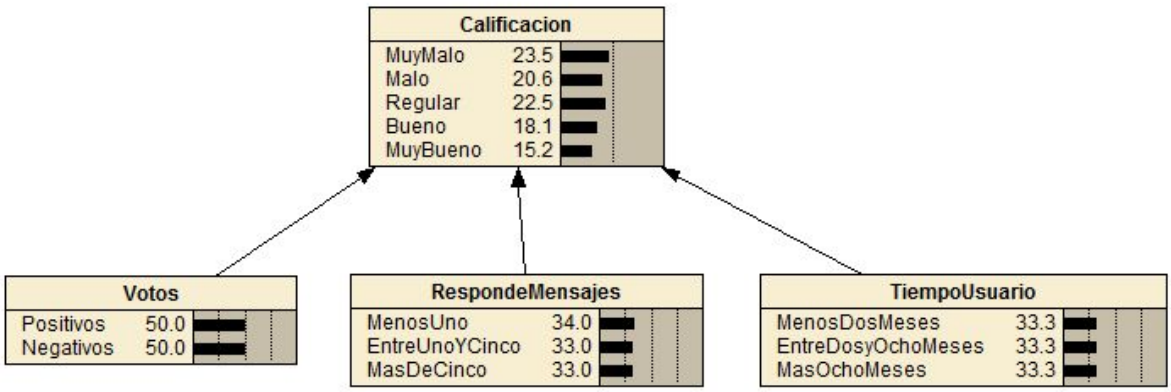

FIGURA 2 - Modelo del clasificador

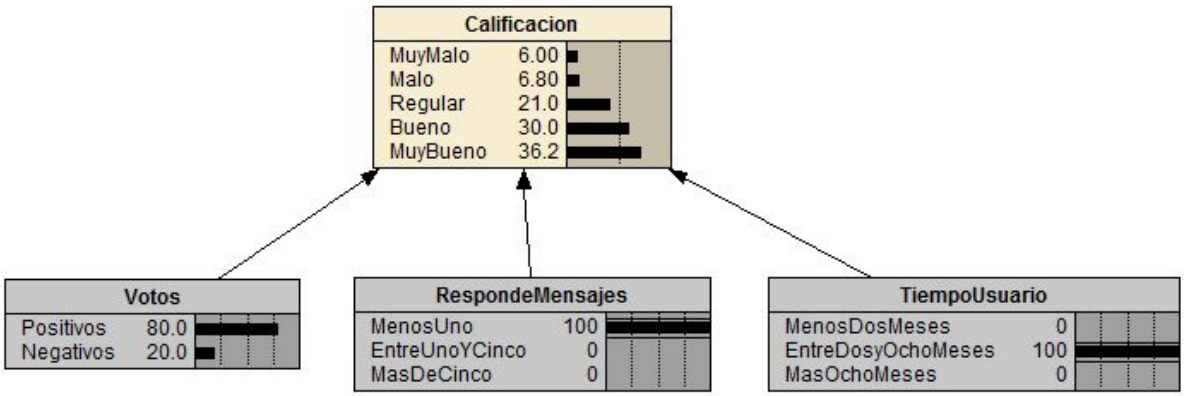

FIGURA 3 - Ejemplo de clasificación 1

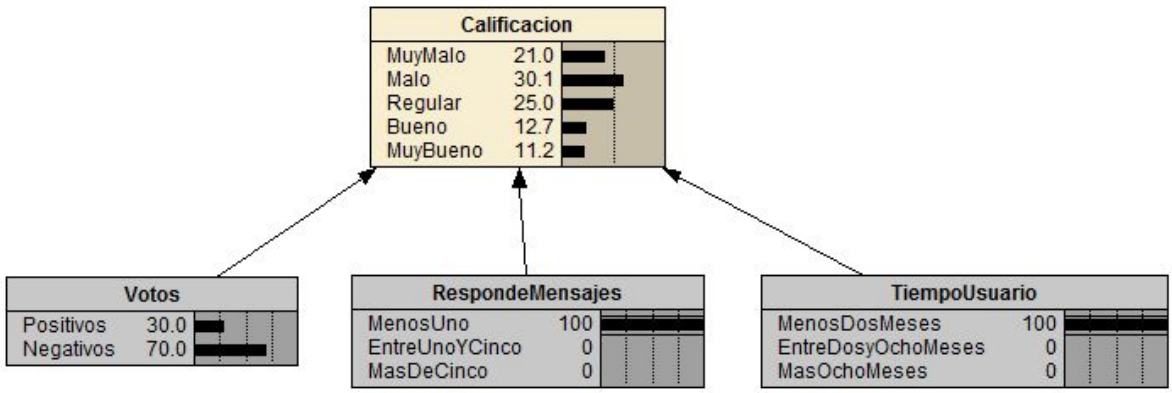

FIGURA 4 - Ejemplo de clasificación 2

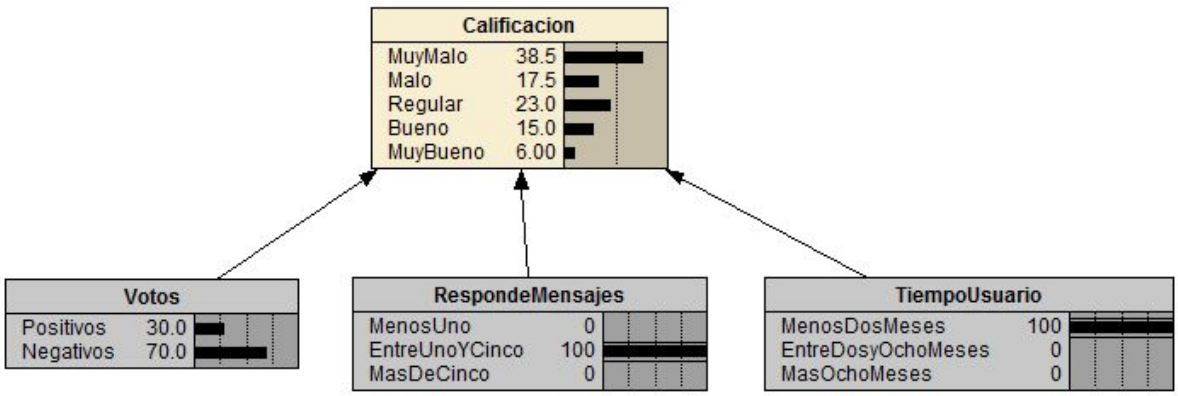

FIGURA 5 - Ejemplo de clasificación 3

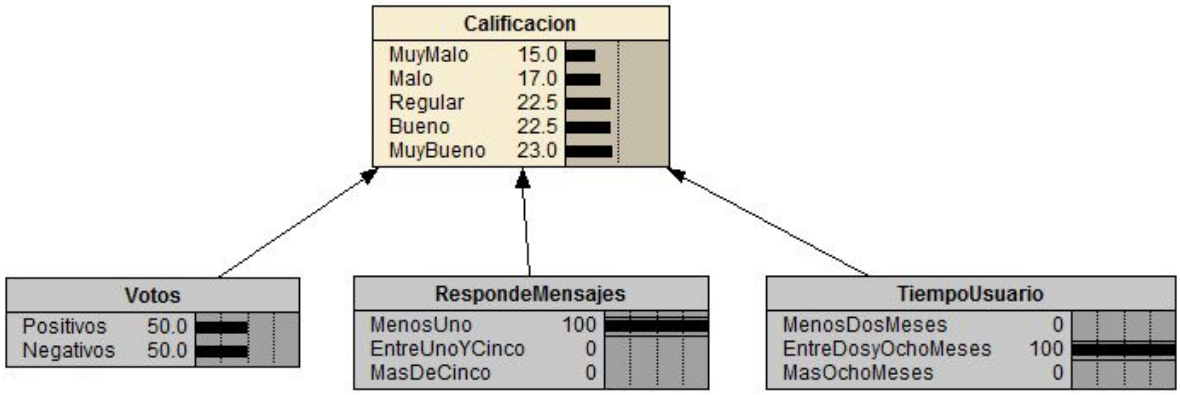

FIGURA 6 - Ejemplo de clasificación 4

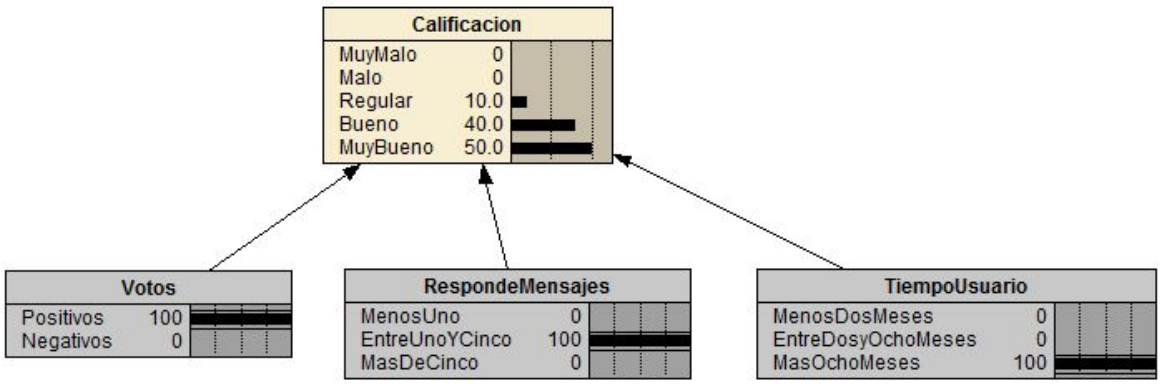

FIGURA 7 - Ejemplo de clasificación 5

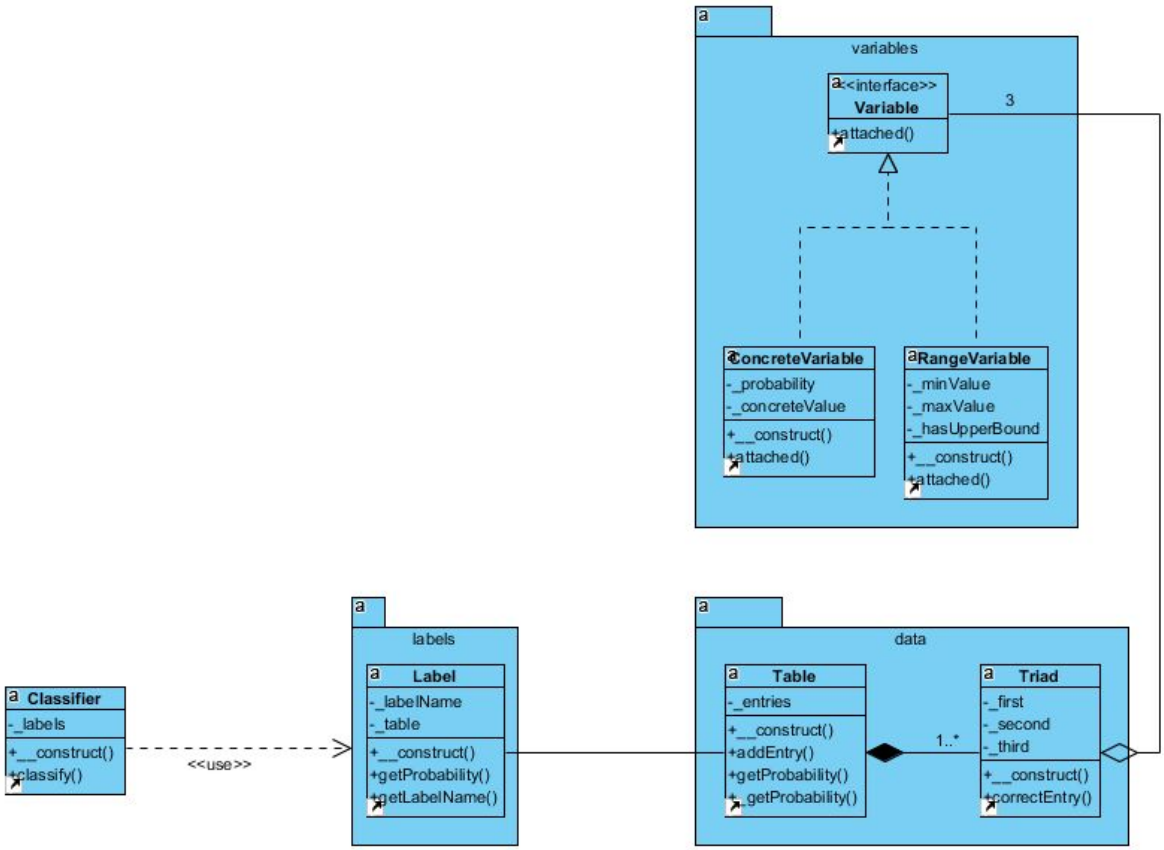

FIGURA 8 - Diagrama de clases - paquetes del sistema.

FIGURA 9 - Interfaz gráfica aplicación que utiliza el clasificador

FIGURA 10 - Ejemplo clasificación en aplicación

Votos	RespondeMensajes	TiempoUsuario	MuyMalo	Malo	Regular	Bueno	MuyBue...
Positivos	MenosUno	MenosDosMeses	0	0	25	40	35
Positivos	MenosUno	EntreDosyOchoMeses	0	0	20	35	45
Positivos	MenosUno	MasOchoMeses	0	0	20	30	50
Positivos	EntreUnoYCinco	MenosDosMeses	0	0	30	50	20
Positivos	EntreUnoYCinco	EntreDosyOchoMeses	0	0	10	50	40
Positivos	EntreUnoYCinco	MasOchoMeses	0	0	10	40	50
Positivos	MasDeCinco	MenosDosMeses	0	30	50	10	10
Positivos	MasDeCinco	EntreDosyOchoMeses	0	20	50	20	10
Positivos	MasDeCinco	MasOchoMeses	0	10	50	30	10
Negativos	MenosUno	MenosDosMeses	30	43	25	1	1
Negativos	MenosUno	EntreDosyOchoMeses	30	34	25	10	1
Negativos	MenosUno	MasOchoMeses	25	29	35	10	1
Negativos	EntreUnoYCinco	MenosDosMeses	55	25	20	0	0
Negativos	EntreUnoYCinco	EntreDosyOchoMeses	45	40	15	0	0
Negativos	EntreUnoYCinco	MasOchoMeses	30	50	20	0	0
Negativos	MasDeCinco	MenosDosMeses	80	20	0	0	0
Negativos	MasDeCinco	EntreDosyOchoMeses	70	30	0	0	0
Negativos	MasDeCinco	MasOchoMeses	60	40	0	0	0

TABLA 1 - Probabilidades condicionales para nodo hipótesis

7. REFERENCIAS

[1]FRIEDMAN M., GEIGER D. Y GOLDSZMIDT M. (1997): Bayesian network classifiers. Kluwer Academic Publishers.Netherlands.

[2]NORSYS SOFTWARE CORP. (1994-2016): Netica Application. Free Version . https://www.norsys.com/netica.html.

[3]NAIVE BAYES CLASSIFIER. (2016): Clasificador de usuarios. https://guarded-fortress-67067.herokuapp.com/.