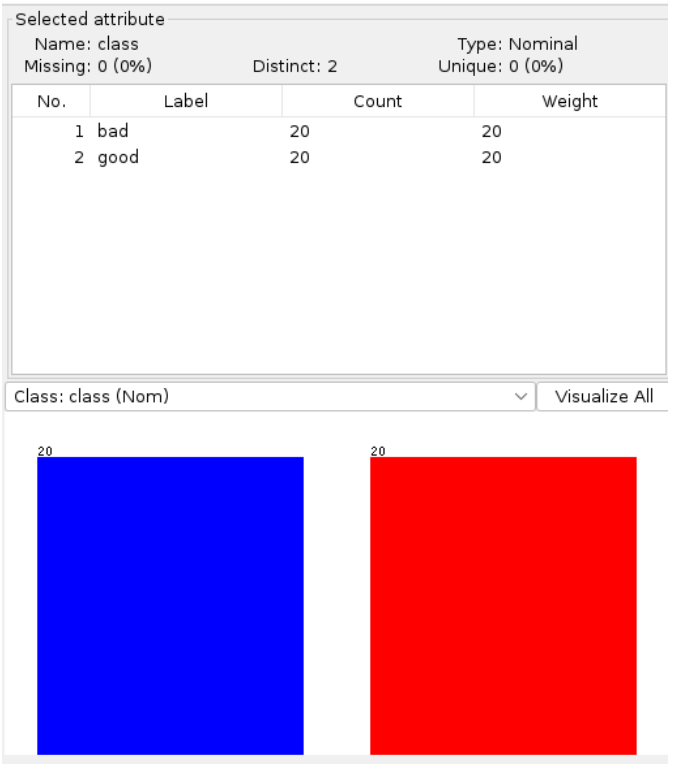
Alberto García Martín

Trabajo 2 - Aprendizaje no supervisado

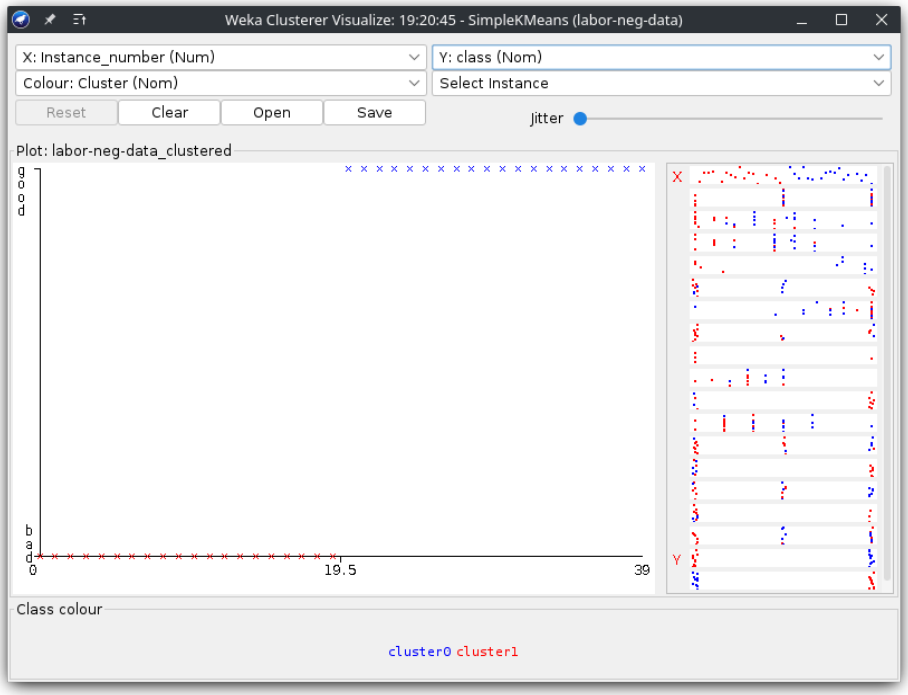
Para este trabajo se ha seleccionado un conjunto de datos que describe la calidad de distintos trabajos en la industria canadiense (labor.arff), basándose en 16 atributos distintos, entre los que se incluyen tanto atributos categóricos como continuos. También contiene un atributo de clase que clasifica las instancias en buenas o malas.

El conjunto de datos contiene 20 instancias clasificadas como malas y 37 instancias clasificadas como buenas, por lo que se han eliminado 17 instancias aleatorias de la clase “good” para balancear el conjunto de datos. Posteriormente, se han ordenado los datos según la clase.



Como modelo de clustering se ha aplicado el algoritmo **k-medias sencillo** (SimpleKMeans en Weka) para que genere 2 clústeres, como función de distancia se ha seleccionado la distancia Manhattan. Después de aplicar el algoritmo se puede destacar que cada centroide de los clústeres tienen atributos de clase distintos, uno “good” y otro “bad”, además el tamaño de ambos clústeres es de 20, por lo que es posible que haya identificado cada clúster con las distintas clases.

Si se visualizan los clústeres poniendo la clase en el eje Y se observa que, efectivamente, cada clúster contiene todas las instancias de una clase. El clúster 0 contiene todas las instancias de la clase “good” y el clúster 1 contiene todas las instancias de la clase “bad”. Como se puede observar en la siguiente figura.



Antes de seguir con la aplicación de un algoritmo de asociación, se ha vuelto a cargar el conjunto de datos, restaurando las instancias borradas, y se ha aplicado un filtro de discretización (unsupervised.attributes.Discretize en Weka). Como parámetros del filtro se ha reducido el número de particiones a 5 y se ha puesto como verdadero el atributo de useEqualFrequency para que las particiones tengan una frecuencia similar.

A continuación, se ha aplicado el algoritmo de asociación **apriori**, para generar las distintas reglas de asociación. Se ha usado CAR para obtener solo reglas que tengan una clase en el consecuente, con el objetivo de encontrar las combinaciones de atributos más relevantes para cada clase.

Las 10 mejores reglas encontradas son las siguientes:

```
Best rules found:

 1. contribution-to-dental-plan=full 13 ==> class=good 13    conf:(1)
 2. wage-increase-first-year='(4.8-inf)' 12 ==> class=good 12    conf:(1)
 3. pension=none 11 ==> class=bad 11    conf:(1)
 4. wage-increase-second-year='(4.2-4.9]' 9 ==> class=good 9    conf:(1)
 5. duration='(1.5-2.5]' contribution-to-dental-plan=full 9 ==> class=good 9    conf:(1)
 6. wage-increase-first-year='(4.15-4.8]' 11 ==> class=good 10    conf:(0.91)
 7. working-hours='(-inf-36.5]' 11 ==> class=good 10    conf:(0.91)
 8. working-hours='(39.5-inf)' vacation=below_average 11 ==> class=bad 10    conf:(0.91)
 9. statutory-holidays='(-inf-10.5]' vacation=below_average 10 ==> class=bad 9    conf:(0.9)
10. statutory-holidays='(10.5-11.5]' longterm-disability-assistance=yes 10 ==> class=good 9    conf:(0.9)
```

Se pueden destacar algunas reglas como la #1 que indica que todas las instancias con una contribución a un plan dental íntegro corresponden a trabajos de calidad. Mientras tanto, la regla #3 indica que todas las instancias sin pensión corresponden a trabajos malos.

También se observa, en la regla #7, que casi todas las instancias que se encuentran en la horquilla más baja de horas trabajadas (<36.5 h) corresponden a buenos trabajos. Mientras que en la regla #8, casi todos los trabajos en la horquilla más alta de horas trabajadas (>39.5 h) y con unas vacaciones por debajo de la media, corresponden a trabajos malos.