732A75 Advanced Data Mining laboratory 3 report

Yuki Washio (timwa902) and Nicolas Taba (nicta839)

March 24, 2021

1 Introduction

The aim of this laboratory exercise is to test the limits of the clustering algorithm using the distance metric and understand that the clustering algorithms can sometimes fail to provide clusters that correspond to classes that we (humans) may intuit better.

In order to illustrate this, we are using the MONK dataset. The dataset is presented on the UCI machine learning repository as well as more in-depth for our particular problem in this article. This dataset presents 124 data points associated to 7 categorical attributes for the purposes of classification. The MONK1 dataset is used to try to classify (class 0 or class 1) using the following constraints: a1 = a2 and a5 = 1, where a1, a2, a5 are attributes of the data points.

We will first start an exploratory analysis and clustering using appropriate clustering methods for our problem using the distance as a metric and trying to improve on it as much as possible before turning to assosciation analysis.

2 Clustering

The dataset presents categorical data. We should use hierarchical clustering methods to create our clusters and obtain the desired 2 clusters.

We are going to use the complete link and kmeans that use euclidean distance as a metric for 2 clusters.

```
=== Model and evaluation on training set ===

Clustered Instances

0       84 ( 68%)
1       40 ( 32%)


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 41 21 | 0
 43 19 | 1

Cluster 0 <-- 1
Cluster 1 <-- 0

Incorrectly clustered instances :       60.0     48.3871 %
```

(a) complete link hierarchical clustering

```
=== Model and evaluation on training set ===

Clustered Instances

0       77 ( 62%)
1       47 ( 38%)


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 40 22 | 0
 37 25 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       59.0     47.5806 %
```

(b) kmeans clustering

Figure 1: Results of clustering algorithm

The clustering algorithm does not perform well and guesses incorrectly about half of the time. This is very bad performance because it is only slightly better than a random guess. This is because although some of the attributes of some data points are similar, they belong to different classes. We cannot distinguish the different class instances using distance as a metric in this case.

We can slightly improve the clustering by creating an outlier cluster. We also use single link hierarchical clustering. This improves the performance of the clustering algorithm, but is still not sufficient enough.

```
=== Model and evaluation on training set ===

Clustered Instances

0       67 ( 54%)
1       56 ( 45%)
2        1 (  1%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 41 21  0 | 0
 26 35  1 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1
Cluster 2 <-- No class

Incorrectly clustered instances :       48.0     38.7097 %
```

Figure 2: Single link hierarchical clustering

The performance improves from random guess to 62% accuracy.

3 Assosciation analysis

We perform association analysis using the apriori algorithm and find up to 19 rules using a minimum support of 0.05. The rules created are presented in Fig.3.

```
Best rules found:

 1. attribute#5=1 29 ==> class=1 29    conf:(1)
 2. attribute#1=3 attribute#2=3 17 ==> class=1 17    conf:(1)
 3. attribute#3=1 attribute#5=1 17 ==> class=1 17    conf:(1)
 4. attribute#5=1 attribute#6=1 16 ==> class=1 16    conf:(1)
 5. attribute#1=2 attribute#2=2 15 ==> class=1 15    conf:(1)
 6. attribute#1=3 attribute#5=1 13 ==> class=1 13    conf:(1)
 7. attribute#5=1 attribute#6=2 13 ==> class=1 13    conf:(1)
 8. attribute#2=3 attribute#5=1 12 ==> class=1 12    conf:(1)
 9. attribute#3=2 attribute#5=1 12 ==> class=1 12    conf:(1)
10. attribute#1=3 attribute#2=3 attribute#6=2 12 ==> class=1 12    conf:(1)
11. attribute#4=1 attribute#5=1 11 ==> class=1 11    conf:(1)
12. attribute#1=2 attribute#5=1 10 ==> class=1 10    conf:(1)
13. attribute#2=2 attribute#5=1 10 ==> class=1 10    conf:(1)
14. attribute#1=1 attribute#2=1 9 ==> class=1 9    conf:(1)
15. attribute#4=2 attribute#5=1 9 ==> class=1 9    conf:(1)
16. attribute#4=3 attribute#5=1 9 ==> class=1 9    conf:(1)
17. attribute#1=2 attribute#2=2 attribute#3=1 9 ==> class=1 9    conf:(1)
18. attribute#1=3 attribute#2=3 attribute#3=1 9 ==> class=1 9    conf:(1)
19. attribute#3=1 attribute#5=1 attribute#6=1 9 ==> class=1 9    conf:(1)
```

Figure 3: 19 best assosciation rules for MONK1 dataset.

Since this is a binary problem, it is sufficient to study rules that predict class 1. We were able to find 4 rules that describe class 1 (the support of these rules covers all the cases of class 1). The rules that describe class 1 are marked in red in Fig.3 and are the same than the expected rules in the description of the dataset in https://www.researchgate.net/publication/2293492_The_MONK's_Problems_A_Performance_Comparison_of_Different_Learning_Algorithms .

4 Discussion

The clustering algorithm has failed to predict when considering the problem of predicting the class. The algorithm might however give us more insights into other aspects of the dataset. The clustering algorithm fails because some of the attributes are not necessary to the classification. Using distance is not a proper measure to establish class differences.

Association analysis and the establishment of rules gives us in this case a more robust way of classifying the data.