The Wilderness Area Data Set: Adapting the Covertype data set for unsupervised learning

Richard Hugh Moulton[1] *and Jakub Zgraja*[2]

[1] `richard.moulton@queensu.ca`
Department of Electrical and Computer Engineering, Queen's University, Kingston ON, Canada

[2] `jakub.zgraja@pwr.edu.pl`
Department of Systems and Computer Networks, Wrocław University of Science and Technology, Wrocław, Poland

January 30, 2019

Benchmark data sets are of vital importance in machine learning research, as indicated by the number of repositories that exist to make them publicly available. Although many of these are usable in the stream mining context as well, it is less obvious which data sets can be used to evaluate data stream clustering algorithms. We note that the classic Covertype data set's size makes it attractive for use in stream mining but unfortunately it is specifically designed for classification. Here we detail the process of transforming the Covertype data set into one amenable for unsupervised learning, which we call the Wilderness Area data set. Our quantitative analysis allows us to conclude that the Wilderness Area data set is more appropriate for unsupervised learning than the original Covertype data set.

Introduction

Benchmark data sets are ubiquitous in the machine learning literature because they offer a method of evaluating algorithms against others in the literature as well as against a fully understood ground truth. This second requirement skews the number of benchmark data sets that are available for different tasks; the UCI Machine Learning Repository[1] contains four times more data sets intended for supervised learning than for unsupervised learning.

Even those data sets that are intended for unsupervised learning may present challenges, given the lack of an agreed upon ground truth against which to assess the eventual clusterings. Supervised learners are given labelled instances for training and can begin to infer connections between attributes and class labels. Unsupervised learners, on the other hand, are left to look at the attributes only and must discover a structure that is internal to the data set.

In this report we describe how we adapt the Covertype data set, a classic benchmark data set for supervised learning, into one that is more appropriate for use with unsupervised learning. We describe this new data set, which we call Wilderness Area, and present the results of quantitative analysis performed to confirm that the Wilderness Area data set presents a reasonable challenge for clustering algorithms.

[1] Lichman, 2013

The Covertype data set

The Covertype dataset was first used in a machine learning context by Blackard and Dean, as part of Blackard's doctoral thesis[2] and then as part of an academic article[3]. Both of the original works use the data set as the basis of a supervised learning task. The data set's instances are drawn from US Forest Service (USFS) Region 2 Resource Information System data and the classifier must predict the type of forest cover present in a 30 x 30 metre cell given the observed geographic information system (GIS) variables.

Since its introduction, it has become a standard benchmark data set in the literature and has been cited by hundreds of papers. The data set is available as raw data from the UCI Machine Learning Repository[4] and in normalized form from the website of Massive Online Analysis (MOA), an open source framework for data stream mining.[5]

Data set information

Instances in the data set are drawn from four different wilderness areas from the Roosevelt National Forest in north Colorado: Rawah, Neota, Comanche Peak and Cache la Poudre. What makes these wilderness areas particularly useful is that they are largely the product of natural process as opposed to human management.[6] As explained in the UCI Machine Learning Repository's description of the data set[7]:

> Neota (area 2) probably has the highest mean elevational value of the 4 wilderness areas. Rawah (area 1) and Comanche Peak (area 3) would have a lower mean elevational value, while Cache la Poudre (area 4) would have the lowest mean elevational value.

These instances are divided into seven types of forest cover type based on the tree species present. These are, in order, spruce/fir, lodgepole pine, Ponderosa pine, cottonwood/willow, aspen, Douglas-fir, and krummholz. Again, from the data set description[8]:

> As for primary major tree species in these areas, Neota would have spruce/fir (type 1), while Rawah and Comanche Peak would probably have lodgepole pine (type 2) as their primary species, followed by spruce/fir and aspen (type 5). Cache la Poudre would tend to have Ponderosa pine (type 3), Douglas-fir (type 6), and cottonwood/willow (type 4).
>
> The Rawah and Comanche Peak areas would tend to be more typical of the overall data set than either the Neota or Cache la Poudre, due to their assortment of tree species and range of predictive variable values (elevation, etc.) Cache la Poudre would probably be more unique

[2] Blackard, 1998

[3] Blackard and Dean, 1999

[4] Lichman, 2013

[5] Bifet et al., 2010

[6] Blackard, 1998

[7] Blackard, 1999

[8] Blackard, 1999

than the others, due to its relatively low elevation range and species composition.

Exploratory analysis was performed using primary component analysis (PCA) in WEKA[9] and non-negative matrix factorization in Matlab (Figure 1). For both methods the data set was reduced to two dimensions to see the strongest separating effects and for ease of visualization. Inspection suggests that the forest cover type attribute does not lend itself to clusters that would be easy to learn.

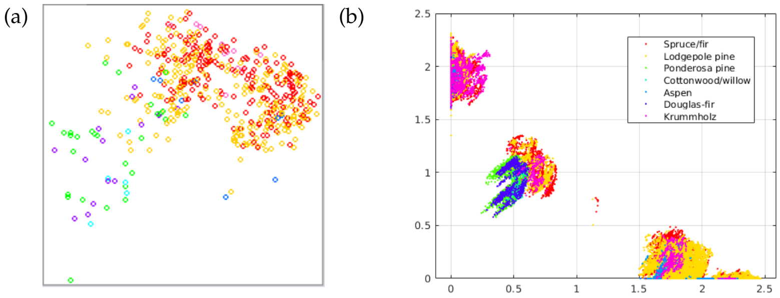

Figure 1: Exploratory analysis of the Covertype dataset —- (a) the first two primary components, (b) non-negative matrix factorization resulting in two factors.

Attribute information

The Covertype data set consists of 54 attributes that are mixed between numerical and binary attributes. An overview of these is given in Table 1.

Blackard gives a full description of why each attribute was included and how it was calculated[10] and we highlight two important observations here. First, elevation is an excellent predictive attribute for determining the forest cover type in a cell because most tree species in the studied wilderness areas grow within specific ranges of altitudes. This is subject to both aspect and slope, which impact both the temperature and available moisture in a given cell. Second, the 40 "Soil Type" attributes are very specific and could be grouped into 11 more general soil classes on the basis of USFS data.

Different subsets of attributes were tested by Blackard and Dean to ensure that each contributed information for the task of predicting forest cover type. Their results showed that classification accuracy was increased for both artificial neural networks and discriminant analysis as the number of attributes was increased.[11]

Summary

The Covertype data set is a very well documented data set that, due to its size, is very desirable to use as a benchmark data set for stream mining tasks. Blackard and Dean, the original authors, validated that

[9] Frank et al., 2016

[10] Blackard, 1998

[11] Blackard and Dean, 1999

Name	Data Type	Description
Elevation	numeric	Elevation in metres
Aspect	numeric	Aspect in degrees azimuth
Slope	numeric	Slope in degrees
Horizontal_Distance _To_Hydrology	numeric	Horizontal distance to the nearest surface water features
Vertical_Distance _To_Hydrology	numeric	Vertical distance to the nearest surface water features
Horizontal_Distance _To_Roadways	numeric	Horizontal distance to the nearest roadway
Hillshade_9am	numeric	Hillshade index at 9 AM, summer solstice
Hillshade_Noon	numeric	Hillshade index at 12 PM (noon), summer solstice
Hillshade_3pm	numeric	Hillshade index at 3 PM, summer solstice
Horizontal_Distance _To_Fire_Points	numeric	Horizontal distance to the nearest wildfire ignition points
Wilderness_Area (4)	binary	Wilderness area designation
Soil_Type (40)	binary	Soil type designation
Cover_Type	nominal	Forest cover type (7)

Table 1: The attributes contained in the Covertype data set (adapted from *Blackard, 1999*)

the attributes included in the data set are useful for the cover type prediction task and the data set has been widely used in the machine learning literature.

That being said, it is clear that this data set is not conducive to clustering. In the next section we address this by transforming the data set into one that can be similarly useful for unsupervised learning.

The Wilderness Area data set

As previous noted, the seven forest cover types included in the Covertype dataset do not describe natural clusters. From Blackard and Dean's description, however, we note that the Wilderness Area attributes have the kind of semantic meaning that we would expect to be able to learn and represent well using a clustering algorithm.[12]

We therefore make use of the normalized data set available from the MOA website and switch the class label in order to support a

[12] Blackard and Dean, 1999

change in task. Now instead of predicting the forest cover type, the objective for the learner is now to cluster instances according to their ground truth Wilderness Area.

Exploratory analysis

We begin by performing the same exploratory analysis as was done for the Covertype data set: PCA using WEKA and non-negative matrix factorization using MATLAB. The results are shown in Figure 2 and appear to indicate a higher degree of separation than was seen with the Cover_Type attribute.

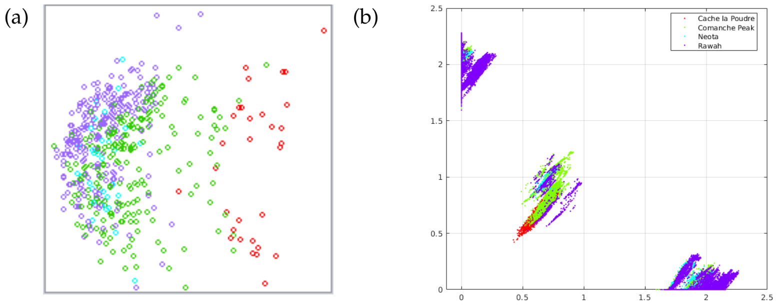

Figure 2: Exploratory analysis of the Wilderness Area dataset — (a) shows the first two primary components, (b) shows non-negative matrix factorization resulting in two factors.

Very noticeable is that the Cache la Poudre wilderness area is clearly the most distinct of the four, as assessed in the description for the Covertype data set.[13]

Dimensionality

One aspect of the Covertype data set that made it very difficult for clustering algorithms is its dimensionality. This leads to the curse of dimensionality, which has negative effects on distance metrics and the clusters that are based on them

To address this issue we make an effort to faithfully represent the data set in as low a dimension as possible. We do this by merging the 40 binary Soil_Type attributes into one nominal Soil_Type attribute and by keeping the Cover_Type attribute as a single nominal attribute rather than splitting it into seven binary attributes as was done for the Wilderness Area attributes in the Covertype data set.

The result of this processing is a twelve-dimensional vector of attributes and a single class attribute. This is fewer than a quarter of the attributes for the Covertype data set without sacrificing any of the GIS data represented. The attributes for the Wilderness Area data set correspond to the rows in Table 1 with nominal attributes replacing binary.

[13] Blackard, 1999

Quantitative Analysis

ALTHOUGH WE MIGHT BE SATISFIED with our intuitions as laid out in the previous section, we have also conducted quantitative analyses to ensure that we have achieved our goal.

Attributes

At the attribute level, we calculate each attribute's Pearson correlation coefficient and Information Gain with respect to its data set's label using Weka.[14] Figure 3 shows both of these measures for the top 12 attributes from the two data sets.

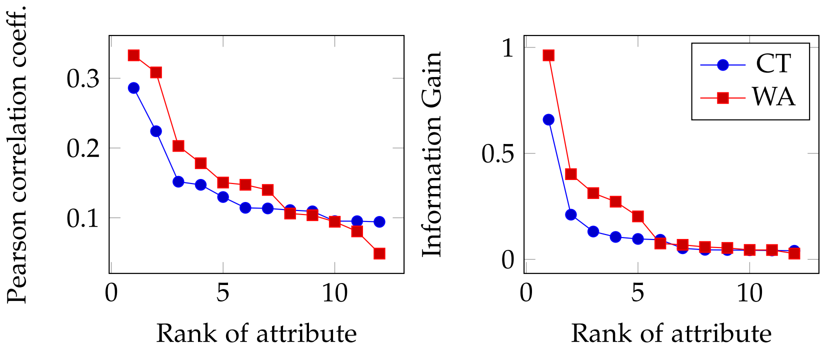

Figure 3: The utility of different attributes in predicting the class label for the Covertype and Wilderness Area data sets.

It is clear from both graphs that transforming the data set has had a positive effect on how well each individual attribute relates to the ground truth labelling: the highest ranking attributes for both measures score higher for the Wilderness Area data set while the remaining attributes generally show fairly even scores.

Clusters

At the cluster level, we measure the silhouette coefficient for both data sets assuming a "perfect" clustering where the ground truth labels are used to indicate cluster membership The silhouette coefficient for a clustering is the mean silhouette value across all instances in the data set and it ranges between -1, which indicates the poorest clustering, and 1, which indicates the best clustering. The silhouette coefficient is useful because it is an internal measure of cluster quality, meaning we can use it to assess how good the clusters represented by a given labelling are. It was used by Kremer et al. as a benchmark measure for their design of a new external measure of cluster quality.[15]

[14] Frank et al., 2016

[15] Kremer et al., 2011

MATLAB was used to calculate the silhouette coefficients for both data sets. For the purposes of computational feasibility, the silhouette coefficient was calculated for the first of ten stratified folds for both of the data sets. These were -0.045 for Covertype and 0.0477 for Wilderness Area (Figure 4). While the difference in values is small, the silhouette coefficient for the Wilderness Area data set is higher and is above 0 – both of which indicate a more natural clustering.

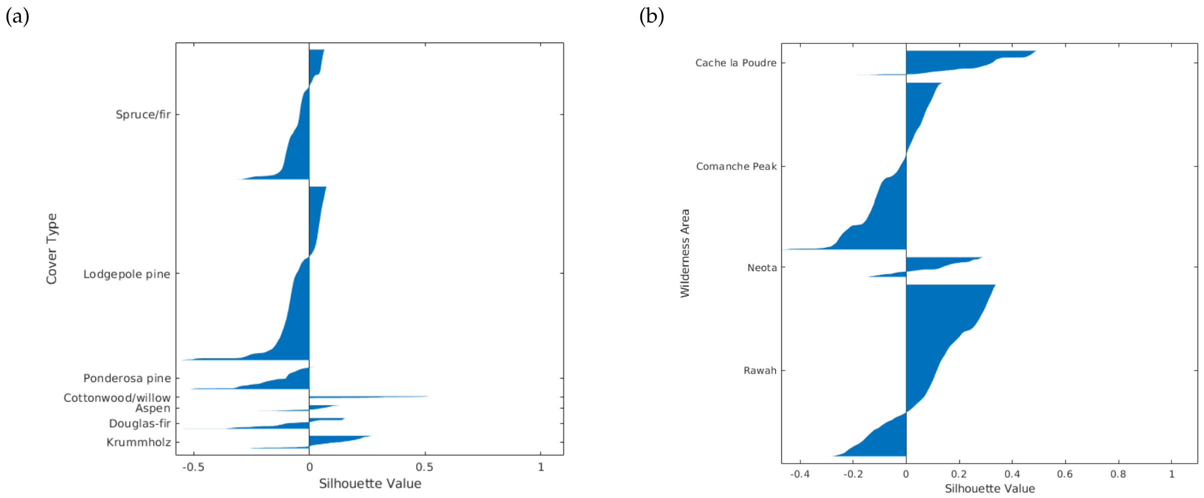

Figure 4: Silhouete analysis — (a) the Covertype data set; and (b) the Wilderness Area data set.

Conclusion

In this report we have detailed the Covertype data set and the reasons why it has become a classic benchmark data set for supervised learning tasks in the machine learning literature. We also noted, however, that it was not well adapted to being used as a benchmark data set for unsupervised learning.

Inspired by the thorough documentation of the data set, we therefore transformed the Covertype data set into the Wilderness Area data set. Using the same domain and semantic meaning, the Wilderness Area data set changes the task from predicting the forest cover type to clustering instances by wilderness area. The quantitative analysis that we performed confirms that, although the clustering task will remain challenging, the Wilderness Area data set is more conducive to finding clusters than the original Covertype data set.

References

Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. Moa: Massive online analysis. *Journal of Machine Learning Research*, 11(May):1601–1604, 2010.

Jock A. Blackard. *Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types*. Doctoral, Colorado State University, 1998.

Jock A. Blackard. The Forest Covertype dataset, 1999. URL `https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.info`.

Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999. ISSN 01681699. DOI: 10.1016/S0168-1699(99)00046-0.

Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, fourth edition, 2016.

Hardy Kremer, Philipp Kranen, Timm Jansen, Thomas Seidl, Albert Bifet, and Geoff Holmes. An Effective Evaluation Measure for Clustering on Evolving Data Streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*, pages 868–876, New York, 2011. ACM Press. ISBN 9781450308137.

M Lichman. UCI Machine Learning Repository, 2013. URL `http://archive.ics.uci.edu/ml`.