Benefits of the decision tree method compared to the Principal Component Analysis method for classifications

Mai Anh Bui[1]
Carnegie Foundation for the Advancement of Teaching

08/12/2022

When solving a classification problem, the decision tree method is among the first, if not the first, idea that a data scientist considers for use as an accurate and powerful classification method. Advanced decision tree models such as the random forest and the extreme gradient boosting (XGBoost) methods are widely used among data scientists (Sarker, 2021). The principal component analysis (PCA) method, on the other hand, is on top of a data scientist's mind to solve a clustering problem or for dimensionality-reduction purposes. This document summarizes key differences between the decision tree and PCA methods and the benefits of using the decision tree method to classify educational institutions compared to the PCA method. The basic decision tree model utilizes one single tree, starts at the root node of the tree, then moves down the tree branches corresponding to the attribute values to classify institutions into different leaves, i.e., class labels or categories. The random forest model utilizes multiple trees simultaneously, takes the majority vote from the outputs of multiple uncorrelated trees to make a collective decision, and outperforms any of the constituent trees. The XGBoost method utilizes multiple trees sequentially, and each subsequent tree improves on the errors of the previous tree. The PCA is a well-known unsupervised learning method and is often used as a dimensionality-reduction technique to reduce the dimensional space in a dataset without losing the variation inherent in it. The comparative strengths of the decision tree method over PCA are outlined below, noting the technical, interpretative, and practical benefits of the decision tree method compared to PCA.

1. **The decision tree detects the most important and interpretable features when given hundreds of input measures**

   The decision tree method does not require the analyst to determine the best input measures; PCA, conversely, requires the analyst to theorize which variables to include and which to exclude. Given hundreds or thousands of input measures, the decision tree method identifies the best features that can partition institutions accurately. Principal components in the PCA method have low interpretability; principal components are composites of the original features and are not as interpretable as the original features. However, the decision tree model gives the importance score for each input measure to notify researchers the most useful features to classify institutions.

2. **There are advanced decision tree models to accompany the basic decision tree model**

   While the PCA method provides one single principal component table to cluster institutions and each institution has one single label result, the random forest and XGBoost methods accompany the basic tree model to give more robust results. When all three tree models determine the same label for an institution, researchers can be confident

[1] I thank Jon Norman and Katharine Bartholomew for their valuable feedback.

about the classification result. In a case where there is difference in the results, researchers can investigate further to determine a final classification decision.

3. **The decision tree method can reproduce classification rules with a high accuracy rate**

   While external researchers need to spend time and resources to understand deeply the principal component table, the cutoff values, and the methodology in the PCA method, the decision tree method with its built-in algorithm defines its own classification rules and can replicate the original classifications with a high accuracy rate. This benefit helps external researchers classify institutions correctly, without the need to understand deeply the original methodology. The three decision tree models (the basic decision tree, the random forest, and the XGBoost) learn from historical 2018 Carnegie Classifications to predict 2021 Carnegie Classifications for R1 and R2 institutions with prediction accuracy rates of above 94 percent. The XGBoost model outperforms with a 95.7 percent prediction accuracy rate.

   The PCA method may not be able to replicate results from other social and economic mobility sources. However, the decision tree method can learn from the Washington Monthly, and CollegeNet existing classifications of institutions' economic and social mobility, which serve as historical information, then apply to our current dataset to provide how Washington Monthly and CollegeNet would classify institutions based on our current dataset. To apply the decision tree method to Washington Monthly and CollegeNet datasets, we need to convert their ranking system into a classification system.

4. **The decision tree method illustrates that an institution can reach its goal through multiple paths**

   The flow-chart of the current Carnegie Classification system illustrates one single path that an institution can reach a particular category or label. However, there is no limit in terms of how many times a leaf (i.e., a label or a category) can appear in different layers and different paths of a decision tree. This means that there are multiple ways that an institution can achieve a label. That is, there are multiple different ways that an institution can achieve the goal.

5. **The decision tree method retains the development over time of educational institutions**

   The PCA method groups unlabeled educational institutions based on their similarities or differences at a given time. The decision tree utilizes historical classifications of educational institutions, trains algorithms, identifies features (i.e., measures) that are most important to demonstrate the existing classifications, and predicts institutions' future outcomes accurately. If the future values of important features can be pre-estimated, the decision tree can predict institutions' future labels. For example, if an institution's endowment is estimated to be increasing next year, and endowment is an important feature, the institution's development in the upcoming year can be predicted.

The decision tree method, like the discriminant analysis method, is a supervised learning technique and requires known precedent upon which to model the data. The decision tree method can build upon historical information in the Carnegie Classifications' basic methodology to predict institutions' future labels. In the first year of the social and economic mobility classifications when there is no historical data to train the model, the decision tree method can serve two purposes. It can be a visualization tool to illustrate classification rules from the root node of the tree, then moves down the tree branches corresponding to the attribute values to classify institutions into different leaves, i.e., class labels or categories. The decision tree can also learn quickly from other existing social and economic mobility ranking and classification systems, such as the Washington Monthly and CollegeNet, to apply to our own data to create classification references for educational institutions.

6. **The decision tree has its own drawbacks.**

The decision tree works better with discrete data than continuous data. This drawback is not relevant for the Carnegie Classifications with discrete labels. The PCA method, on the other hand, is recommended for continuous data; the use of dummy variables in the PCA is not justified, as PCA "as is" is only suitable for continuous data (Kolenikov and Angeles, 2004). Another drawback of the decision tree method is that it can create different trees when we run the model multiple times using the same historical data; the prediction accuracy rates can slightly change in different versions. We solve the problem by utilizing other advanced decision tree models to accompany the basic tree model such as the random forest, or the XGBoost methods, to obtain robust results.

One of the earliest and most influential papers for the construction of socio-economic indices that used PCA was Filmer and Pritchett (2001) (Kolenikov and Angeles, 2004). This PCA method faced a critique that the PCA "as is" is only suitable for continuous data. The decision tree method is a popular method to solve a classification problem and works well with discrete data. In this memo, I also outlined other drawbacks of the PCA method such as the low interpretability of the principal components, its inadaptability to learn from historical information and other resources, and its one single route to success for educational institutions. The decision tree method solves these problems by providing important scores for each input measure to notify researchers the most useful measures to classify institutions, retaining institutions' historical information for the future prediction, learning from, and replicating other classification systems without the need for deep understanding of the original methodologies, and providing institutions with multiple paths to success.

Appendix tables

Table 1: The comparison of the decision tree and the PCA methods [2]

Decision tree	PCA
Supervised learning: The algorithm uses **labeled** datasets to classify data or predict outcomes accurately	**Unsupervised learning:** The algorithm discovers hidden patterns or data groupings without the need for human intervention, analyzes and clusters **unlabeled** datasets
A classification method: The method draws conclusions on how entities should be labeled or defined, uses an algorithm to accurately assign test data into specific categories	**A clustering method:** a data mining technique which groups unlabeled data based on their similarities or differences

Table 2: Various types of machine learning techniques (Sarker, 2021)

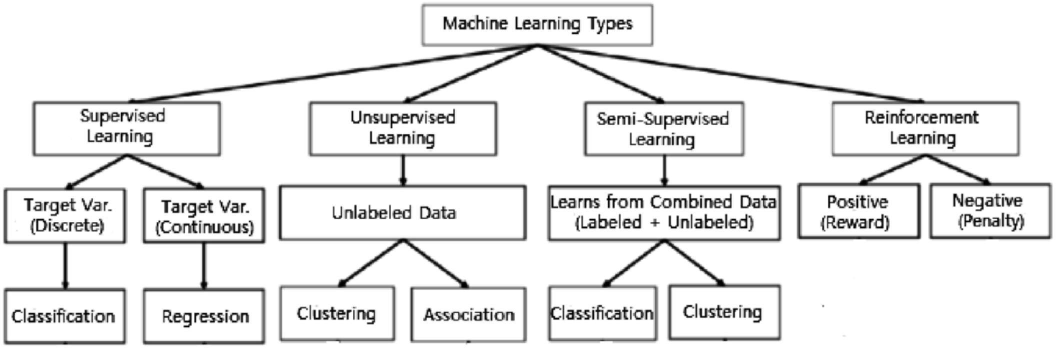
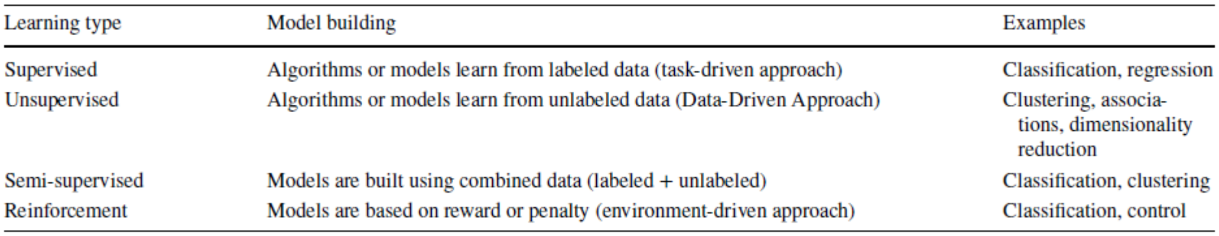


Table 3: Various types of machine learning techniques with examples (Sarker, 2021)



Learning type	Model building	Examples
Supervised	Algorithms or models learn from labeled data (task-driven approach)	Classification, regression
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, associations, dimensionality reduction
Semi-supervised	Models are built using combined data (labeled + unlabeled)	Classification, clustering
Reinforcement	Models are based on reward or penalty (environment-driven approach)	Classification, control

References:

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Kolenikov, S., Angeles, G. (2004), The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices, Working paper WP-04-85, MEASURE/Evaluation project, Carolina Population Center, University of North Carolina, Chapel Hill.

[2] For more detailed information, please refer to IBM Cloud Education and “An introduction to statistical learning” by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN computer science*, *2*(3), 160.