May 26, 2017 Connor George

Introduction

All algorithms that are used for data mining or machine learning have advantages and disadvantages. Typically the engineer or scientist performing data analysis will choose an algorithm that is well suited for some particular application. However, it is often a good choice to use multiple algorithms together to get better performance than any individual algorithm could achieve on its own. An algorithm that uses this technique and combines multiple other machine learning algorithms is referred to as an ensemble algorithm.

Ensemble algorithms have many advantages. They are often able to cancel out weaknesses of their sub-algorithms since a weakness in one can be overcome by the others. This tends to lead to a more robust and accurate algorithm. However, a downside of ensemble algorithms is that they are inherently more complicated than an individual algorithm, and therefore they can require more tuning.

Ensemble algorithms work by combining the predictions of their sub-algorithms into a single overall prediction. For this reason, it is important that the sub-predictions are combined in a proper way, and that they accurately reflect the intuition of the predictions from the sub-algorithms. This project will focus on the manner in which those predictions are combined into a final prediction.

Project

For this project, a comparison will be made between methods of combining the predictions of sub-algorithms in an ensemble algorithm. For clarity, I will define several terms that will be used for the rest of this write-up. The term "interior algorithm" or "sub-algorithm" will refer to one of the algorithms that make up the ensemble algorithm by making predictions on the data. The term "combination strategy" will refer to the strategy or algorithm that is used to combine the predictions of the interior algorithms. "Interior predictions" and "final predictions" will refer to the predictions made by the interior algorithms and the combination strategy, respectively.

This project will explore how different combination strategies can effect the final predictions made by an ensemble algorithm. Three different combination strategies are being considered for the project. The first is a simple naive method of combination, which is to simply count the interior predictions and make the final prediction based on the most common one. This strategy will serve as a baseline for performance since it is the simplest method.

The next strategy that will be used is a linear classification algorithm that uses the interior predictions of the algorithm as training data. The current plan is to use a decision tree, but this choice will be explored further before work on the project begins. It is hypothesized that the decision tree will outperform the naive counting method. This is because the decision tree will be able to consider relations between the interior predictions. For example, it is possible that some particular pair of the interior algorithms will tend to disagree on certain types of examples. The decision tree could learn to prioritize one of their predictions over the other, or even prioritize some third algorithm that proves to be more accurate.

The final strategy for combining the data will be a non-linear classification algorithm. A neural network will most likely be used for this. It is hypothesized that this combination strategy

will perform best of the three. This is because the kinds of patterns that could show up in the accuracy of the other interior algorithms could be very complicated or subtle. Since the problem is complex, it is theorized that a complex model such as a neural network could yield better results than more simple models.

It is also possible that different strategies will work better in some situations compared to others. It would be undesirable for a result to only reflect better performance in certain unique situations, so the strategies will be rigorously tested in a variety of different ensemble configurations. This will ensure that not only are the results of the experiment more accurate on the particular dataset that is used, but they are also more generalizable to other datasets.

Methods

To test the three different combination strategies, an ensemble algorithm will be written. I plan to write the algorithm in Python, and make use of the scikit-learn package for the implementation of the interior algorithms and the combination algorithm. The exact layout of the algorithm will be altered in order to test the hypotheses in different environments.

One test that will be performed is how the accuracy is affected by different types of interior data mining algorithms. Three layouts will be tested for each of the combination strategies. One layout will be a uniform collection of interior algorithms. For example, all of the interior algorithms could be decision trees (with variation in their parameters).

Another layout will be a uniform distribution of algorithms. For example, if five different algorithms were being used, each type would make up exactly one fifth of all the interior algorithms. Finally, a random distribution of algorithms will be tested. In this, algorithms will be selected at

random and added to the pool of interior algorithms.

Each of these will be tested so that the results of the combination strategy testing are not corrupted by the layout of the interior algorithms. If, for example, the neural network yields the highest accuracy in all three layouts that would be much stronger evidence that the hypothesis is correct than if it only yielded higher performance in certain situations.

Each algorithm, including both the interior algorithms and the combination algorithm will be tuned to provide results which accurately reflect the state of the problem. Each type of data mining algorithm used for interior predictions will first be tuned on the dataset as an individual algorithm. This will ensure that the interior algorithms in the ensemble system are making accurate predictions themselves. This is necessary because the results of the experiment could be corrupted if some of the interior algorithms are working well and some are not. Proper tuning of all algorithms reduces the threats to validity for this experiment.

The performance of each strategy will be measured using several different metrics. The overall accuracy of the system will be the metric that is most central to the experiment. Also, the frequency of false-positives and false-negatives will be considered. Finally, the running time of each algorithm will be measured to determine any possible tradeoffs of the system (i.e. higher accuracy comes at the cost of longer running times).

Data

A dataset has not yet been chosen for this experiment. However, the dataset itself is not the primary focus of this experiment. When choosing the dataset, the primary concern will be to find a dataset that is both representative of a "typical" dataset, and is compatible with a variety of data

mining algorithms. Certain types of data work better with some algorithms than with others. For example, decision trees are not able to interpret numeric data without binning it into categories. In order to minimize the threats to validity present in the experiment, the minimal amount of modification to the data (such as binning) will be performed, and datasets that lend themselves well to this goal will be prioritized.

In order to further reduce threats to validity, the experiment will be performed on multiple datasets if time permits. An inherent principle of a data mining algorithm should hold true across multiple datasets. By testing over multiple datasets, it will be less likely that erroneous results will be found due to irregularities within a single chosen dataset.

Conclusion

This experiment will test various combination methods in ensemble algorithms across different interior algorithm layouts and different datasets. Ideally, a pattern will emerge that demonstrates superiority of one method over the others in some or all situations. However, the experiment will be considered successful if results are found in a scientific way with as few threats to validity as possible. Ensemble algorithms are among the most powerful data mining algorithms today, and insight into the most effective combination strategies could lead to even better performance in real-world data mining applications.

References

- Hamza, Mounir, and Denis Larocque. "An empirical comparison of ensemble methods based on classification trees." *Journal of Statistical Computation and Simulation* 75.8 (2005): 629-643.

- Opitz, David, and Richard Maclin. "Popular ensemble methods: An empirical study." *Journal of Artificial Intelligence Research* 11 (1999): 169-198.

- Dietterich, Thomas G. "Ensemble methods in machine learning." *International workshop on multiple classifier systems*. Springer Berlin Heidelberg, 2000.