Principle Component Analysis for Feature Reduction and Data Preprocessing in Data Science

Hayden Wimmer
Department of Information Technology
Georgia Southern University
hwimmer@georgiasouthern.edu

Loreen Powell
Department of Innovation, Technology, and Supply Chain Management
Bloomsburg University
lpowell@bloomu.edu

Abstract

Medical datasets are large and complex. Due to the number of variables contained within medical data, machine learning algorithms may not be able to induct patterns from the data or may over fit the learned model to the data thereby reducing the generalizability of the model. Feature reduction seeks to limit the number of variables as input by establishing correlations between variables and reducing the overall feature set to the minimum number of possible variables to describe the data. This research seeks to examine the effects of principal component analysis for feature reduction when applied to decision trees. Results indicate that principle component analysis (PCA) may be employed to reduce the number of features; however, the results suffer minor degradation.

Keywords: Feature Reduction, Principal Component Analysis, Medical Data, PCA.

1. INTRODUCTION

Health Information Technology (HIT) is an important topic facing Healthcare facilities and professionals around the world. Specifically, HIT in the form of Electronic Health Records (EHRs) and various electronic medical database systems have the ability to aid and transform traditional ways on the healthcare system by improving the quality of medical care and reducing the cost of the medical care (Fabbri, LeFevre, & Hanauer, 2011). EHRs provide extensive amounts of structured data when data is specifically entered into required fields and unstructured data when data is entered as comments and notes or non-labeled fields. Today, with health paper-based health records being converted to EHRs, the data tends to be structured. It is the migration and the transferring of data in the medical data systems that provides researchers with the best opportunity to use data-mining methods for predictive analysis (Park & Ghosh, 2011).

There are many dimensions to any patient. Some dimensions, such as blood pressure and heart rate, are valid in most medical scenarios. Demographic data adds another set of dimensions to a patient. Furthermore, each specific disease and diagnosis has specific dimensions (e.g. tumor size, type, location in cancer patients). A heart patient will have data specific to heart conditions and a cancer patient data specific to cancer with overlapping features such as vital signs and demographics.

As medical facilities continue to integrate and advances in storage and health information technology progresses, the dimensions for a patient subsequently increase. This added data

provides immense opportunity to discover vital information contained within that can prevent or cure diseases and improve a patient's quality of life.

Considering the number of possible conditions with specific data and features, the number of dimensions that are possible for an individual patient presents challenges for data scientists who aim to perform knowledge discovery and data mining. A dataset with high dimensionality may not be minable causing machine learning algorithms to over fit data or generate incomprehensible rules. Oftentimes, underlying relationships, such as correlation, that can be used to reduce the number of features can provide respite. If two features are highly correlated, one feature can be removed since it can be predicted based on the remaining feature. This work seeks to perform dimensionality reduction on a high feature medical dataset using principle component analysis. This works demonstrates that following PCA, a machine learning algorithm, C4.5, produces a more understandable decision tree. The structure of this work is as follows: section 2 discusses background information, section 3 contains the experimental setup, section 4 presents the results, and section 5 contains conclusions and future directions.

2. BACKGROUND

Dimension Reduction

Dimension reduction is an algorithm design tool used for a multitude of related fields (BARTAL, GOTTLIEB, & NEIMAN, 2014). It specifics the plotting of points in high-dimensional properties to low- dimensionality properties and maintaining some points from the original properties (BRINKMAN & CHARIKAR, 2005). Dimension reduction is the process of removing the number of variables in a data set (ROWEIS & SAUL, 2000). The process is often based upon the correlation among variables. For example, if A and B are correlated at 100% then only 1 of the variables is required for machine learning since we may assume that a implies b and b implies a. C4.5 is a machine learning algorithm for classifying data into tree structures (QUINLAN, 1993). For many years researchers have utilized dimension reduction when searching for nearest and clustering of dimensional points (BRINKMAN & CHARIKAR, 2005).

Principal Component Analysis

PCA is a multivariate technique which extracts important information from data and represents it as a new set of variables called principle components (Abdi & Williams, 2010). PCA is a type of factor analysis that is often employed for dimension reduction in a dataset. PCA is often found in research regarding "data mining, pattern recognition and information retrieval for unsupervised dimensionality reduction" (Omucheni, Kaduki, Bulimo, & Angeyo, 2014). Additionally, (Omucheni et al., 2014) utilized PCA in the processing of patient blood smear images to identify Plasmodium parasites for malaria. The results were successful and provide a foundation for further exploratory work in using PAC techniques within medical data sets.

Machine Learning

Machine learning (ML) involves the automated learning of patterns from data or employing past experiences and data to solve a given problem (Alpaydin, 2014). More specifically, machine learning involves learning structure from examples and is the basis for data mining (Carbonell, Michalski, & Mitchell, 1983). Machine learning can be applied to decision tree induction, neural network, Bayesian classifiers, and association rule mining to name a few examples. In machine learning from data, a data set is broken into a training set and a testing set. The training set is input into the ML algorithm where patterns or models are formed then the models applied to the test dataset to determine accuracy and error rate using common measurements such as classification accuracy, confusion matrices, and ROC curves.

Decision Trees

Decision trees are a type of directed graph which begins with a root node. The root node branches to other nodes in the tree. Nodes are connected in a parent child relationship by an edge. A terminating node is referred to as a leaf node. Decision tree induction is the process of learning decision trees from data. Decision trees are one popular techniques in data mining (Ferreira, 2006) and many common decision tree learning algorithms are based on the work of (Quinlan, 1986) where the ID3 algorithm is introduced as a recursive algorithm using information gain to determine when to divide attributes of a dataset in a parent child relationship. This work has been generalized by (Cheng, Fayyad, Irani, & Qian, 1988) and extended by (Quinlan, 1993) into the C4.5 algorithm and (Quinlan, 2012) as the C5.0 algorithm. While ID3 and C4.5 are open source, C5.0 is a commercial version of the aforementioned decision tree algorithms.

3. EXPERIMENT SETUP

The purpose of this applied research is to begin an examination of the effectiveness of PCA for preprocessing large feature medical data for machine learning purposes. A medical dataset with 88 dimensions from a regional health provider was selected. The medical dataset was structured in CSV format, all attributes as numeric values, and with the first row containing column names. The data were general heterogeneous patient records and were not utilized to treat any disease or treatment. The structured medical data set used was targeted toward determining the possibility of developing a certain condition with each attribute leading to a target for classification purposes. Data attributes included demographic information such as gender, race, and age paired as well as information on smoking habits, blood pressure at intake and discharge, asthma status, etc. Due to the sensitive nature of this data and IRB requirements, data columns and values are masked in the resulting analysis. PCA was performed using JMP by SAS.

As illustrated in figure 1, three paths were taken. The first performs C4.5 against the full dataset. The second uses PCA for dimension reduction and uses variables from the first principle component as input to C4.5. The third performs dimension reduction to the first and second principle component. The Dimension reduction was performed using PCA selecting the important variables. Figure 3 shows the results of the first principle component (PCA1) and the second Principle component (PCA2) screen plot. Initially, the first principle component was selected because it accounted for the greatest possible variance within the data set.

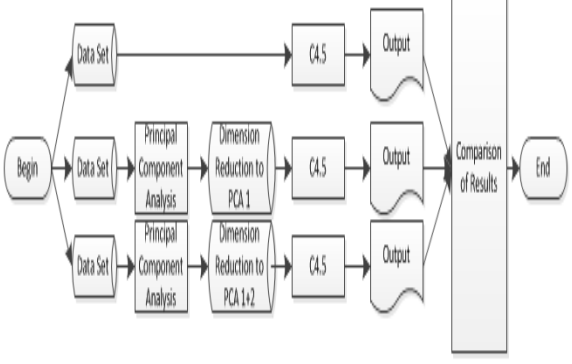

Figure 1: Flow of Experiment

The variables from the first principle component were input to a C4.5 machine learning algorithm for classification. Decision Trees are more easily understood than other machine learning algorithms, such as neural networks; therefore, the C4.5 machine learning algorithm was selected as a test case for PCA in dimension reduction of medical data. Next, for comparison purposes, the variables form PCA1 and PCA2 were selected. The variables for PCA1 and PCA2 were placed into a C4.5 machine learning algorithm for classification. The output was analyzed and compared with the results for only PCA1.

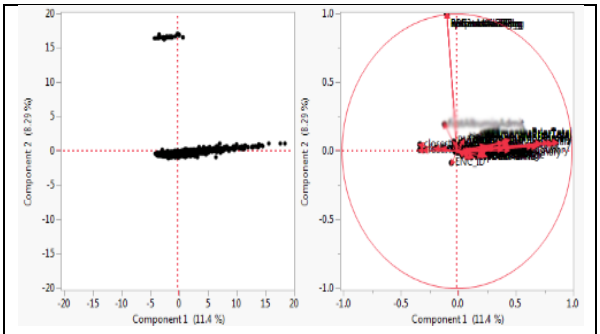

Figure 2: Principle Component 1 and 2 screen plot

4. RESULTS

Preliminary results indicate mixed results on the effectiveness of PCA when dealing with high-dimension medical datasets. Figures 3 and 4 show the results of applying the C4.5 decision tree algorithm to the initial medical data set prior to any feature reduction. The phase performed no dimension reduction with an 81.97% classification accuracy and a 0.566 ROC area.

```
Size of the tree :      167


Time taken to build model: 0.94 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3456               81.9734 %
Incorrectly Classified Instances       760               18.0266 %
Kappa statistic                          0.1275
Mean absolute error                      0.2537
Root mean squared error                  0.393
Relative absolute error                 92.2159 %
Root relative squared error            105.972  %
Coverage of cases (0.95 level)          96.0152 %
Mean rel. region size (0.95 level)      90.9749 %
Total Number of Instances             4216

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.143    0.047      0.375      0.143     0.207      0.147     0.566     0.227     Yes
                 0.953    0.857      0.849      0.953     0.898      0.147     0.566     0.852     No
Weighted Avg.    0.820    0.724      0.771      0.820     0.784      0.147     0.566     0.749

=== Confusion Matrix ===

    a    b   <-- classified as
   99  595 |    a = Yes
  165 3357 |    b = No
```

Figure 3 – Results Prior to Feature Reduction

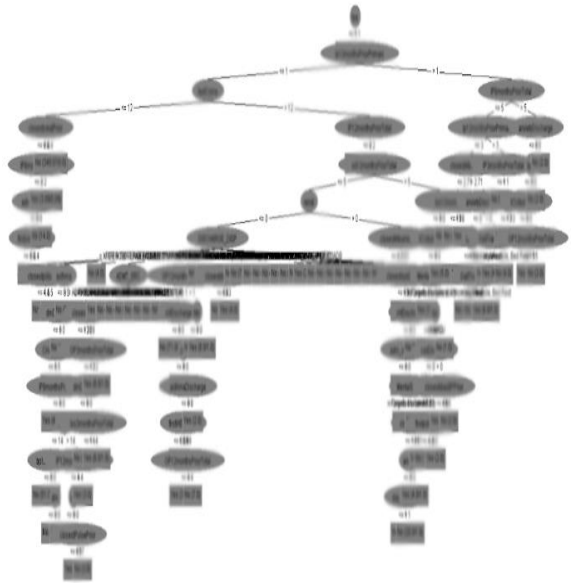

Figure 4 – Decision Tree Prior to Feature Reduction

Next, upon performing dimension reduction using PCA1, the results show in an increase of classification accuracy to 83.56. However, there is also a reduction in the ROC area to 0.543. Please reference Figures 5 and 6 for illustrate results.

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3523               83.5626 %
Incorrectly Classified Instances       693               16.4374 %
Kappa statistic                          0.0739
Mean absolute error                      0.2644
Root mean squared error                  0.3719
Relative absolute error                 96.107  %
Root relative squared error            100.2876 %
Coverage of cases (0.95 level)          99.2884 %
Mean rel. region size (0.95 level)      99.1017 %
Total Number of Instances             4216

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.059    0.011      0.506      0.059     0.106      0.129     0.543     0.221     Yes
                 0.989    0.941      0.842      0.989     0.909      0.129     0.543     0.843     No
Weighted Avg.    0.836    0.788      0.787      0.836     0.777      0.129     0.543     0.740

=== Confusion Matrix ===

    a    b   <-- classified as
   41  653 |    a = Yes
   40 3482 |    b = No
```

Figure 5 – Results After Feature Reduction PCA 1

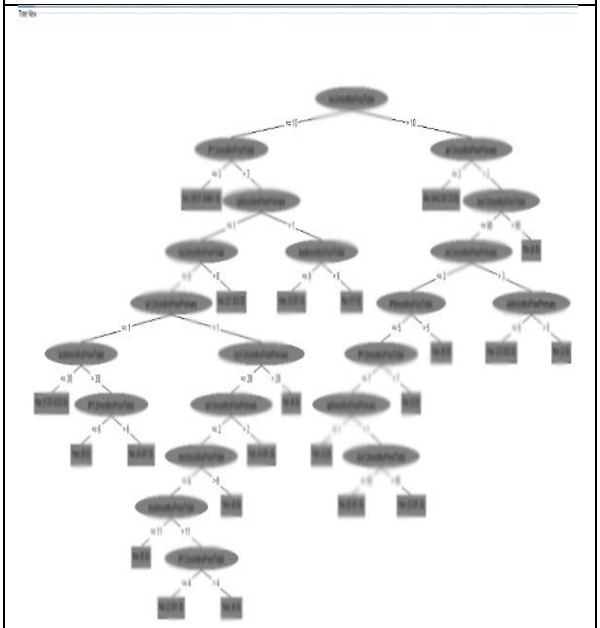

Figure 6 – Decision Tree After Feature Reduction PCA 1

Finally, when reducing dimensions to PCA 1 and PCA2, the results indicated the same classification accuracy as the first PC only of 83.56. Additionally, the ROC was further diminished to 0.513. Please reference Figure 7 and 8 for illustrated results.

```
Incorrectly Classified Instances       693               16.4374 %
Kappa statistic                          0.0501
Mean absolute error                      0.2698
Root mean squared error                  0.3718
Relative absolute error                 98.0479 %
Root relative squared error            100.2684 %
Coverage of cases (0.95 level)          99.5968 %
Mean rel. region size (0.95 level)      99.5849 %
Total Number of Instances             4216

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.039     0.007     0.509      0.039     0.072     0.105     0.513     0.185     Yes
               0.993     0.961     0.840      0.993     0.910     0.105     0.513     0.837     No
Weighted Avg.  0.836     0.804     0.785      0.836     0.772     0.105     0.513     0.729

=== Confusion Matrix ===

   a   b   <-- classified as
  27  667 |   a = Yes
  26 3496 |   b = No
```

Figure 7 – Results After Feature Reduction PCA 1+2

Figure 8 – Decision Tree After Feature Reduction PCA 1+2

One interesting note was the size of the initial tree in Figure 3 was 167 which had more nodes than Figures 5 or 7. This may be explained as there are less features from which to generate a decision tree; however, such a large tree may be over-fit and therefore not generalizable. The classification accuracy of the resulting C4.5 decision trees increases from 81.97% to 83.56%; however, conversely the ROC decreases from 0.57 to 0.55 and 0.51.

While interesting, the mixed results require additional work to fully map the potential of PCA for dimensionality reduction in high-dimension medical data. With medical data, resulting knowledge structures (i.e. decision trees) and the variables in each principle component must be verified by domain experts, such as physicians. It would be necessary for each application of dimension reduction to determine acceptable ranges for diminished results such as classification accuracy and ROC area. In the scenario examined in this work, the tree resulting from PCA 1 is the simplest in terms of structure and human understandability; therefore, a the reduction in ROC area may be an acceptable concession.

5. CONCLUSIONS AND FUTURE DIRECTIONS

The results provided in this work further expand the understanding and effectiveness of using PCA techniques in medical data sets for dimension reduction. The experiments demonstrated that applying PCA prior to decision tree induction has mixed results, namely increasing classification accuracy but decreasing ROC area. One notable result was the simplification of the resulting decision trees after the application of PCA. Human understandability and generalizability are important characteristics of decision trees; therefore, the concession may be worthwhile. The decision tree from the full dataset contained 167 nodes thereby demonstrating the possibility of over-fitting and a lack generalizability. It is noted that determining acceptable parameters for changes in classification accuracy and ROC area are application specific and require domain expertise for appropriate judgement. This research is not without limitations as it is limited by a single medical data set, only reviews one method of feature reduction, and one machine learning algorithm. Future research will address the aforementioned limitations. Implications of this research include providing data scientists and practitioners a first step when dealing with high-feature medical datasets and provides a direction for future development and application of dimension reduction in clinical informatics.

6. REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433-459.

Alpaydin, E. (2014). *Introduction to machine learning*: MIT press.

Bartal, Y., Gottlieb, L.-A., & Neiman, O. (2014). *On the Impossibility of Dimension Reduction for Doubling Subsets of ℓp.* Paper presented at the Annual Symposium on Computational Geometry.

Brinkman, B., & Charikar, M. (2005). On the impossibility of dimension reduction in l 1. *Journal of the ACM (JACM), 52*(5), 766-788.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An overview of machine learning *Machine Learning* (pp. 3-23): Springer.

Cheng, J., Fayyad, U. M., Irani, K. B., & Qian, Z. (1988). *Improved Decision Trees: A Generalized Version of ID3.* Paper presented at the ML.

Fabbri, D., LeFevre, K., & Hanauer, D. A. (2011). *Explaining accesses to electronic health records.* Paper presented at the Proceedings of the 2011 workshop on Data mining for medicine and healthcare.

Ferreira, C. (2006). Decision Tree Induction *Gene Expression Programming* (pp. 337-380): Springer.

Omucheni, D. L., Kaduki, K. A., Bulimo, W. D., & Angeyo, H. A. (2014). Application of principal component analysis to multispectral-multimodal optical image analysis for malaria diagnostics. *Malaria journal, 13*(1), 485.

Park, Y., & Ghosh, J. (2011). *A generative framework for predictive modeling using variably aggregated, multi-source healthcare data.* Paper presented at the Proceedings of the 2011 workshop on Data mining for medicine and healthcare.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81-106.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning* (Vol. 1): Morgan kaufmann.

Quinlan, J. R. (2012). C5.0: An Informal Tutorial.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(5500), 2323-2326.