Using Random Projections to Identify Class-Separating Variables in High-Dimensional Spaces

Anushka Anand*, Leland Wilkinson† and Tuan Nhon Dang‡
Department of Computer Science, University of Illinois at Chicago

ABSTRACT

Projection Pursuit has been an effective method for finding interesting low-dimensional (usually 2D) projections in multidimensional spaces. Unfortunately, projection pursuit is not scalable to high-dimensional spaces. We introduce a novel method for approximating the results of projection pursuit to find class-separating views by using random projections. We build an analytic visualization platform based on this algorithm that is scalable to extremely large problems. Then, we discuss its extension to the recognition of other noteworthy configurations in high-dimensional spaces.

Index Terms: H.5.2 [User Interfaces]: Graphical user interfaces (GUI)— [H.2.8]: Database Applications—Data Mining

1 INTRODUCTION

Friedman and Tukey [4] introduced a method, called Projection Pursuit (PP), for identifying noteworthy 2D projections of points in a multivariate space. Unfortunately, PP is not a scalable algorithm. For high-dimensional applications, PP can fail to converge in practical time. This difficulty led us to consider how to approximate the results of PP with a scalable algorithm. We leverage the Johnson-Lindenstrauss theorem [6] using unit-weighted random projections [1]. By working with point-to-point distances in low-dimensional, *non-axis-parallel* subspaces, we escape the curse of dimensionality.

The main use of the random projection theorem in data mining has been to perform preliminary random projection in order to reduce the dimensionality of a space to a manageable subspace before applying classification algorithms such as Support Vector Machines or decision trees that are not readily scalable to extremely high-dimensional spaces. Our approach was different. We developed a visual-analytic classifier based on iterated random projections [11] that approximates the distribution-free flexibility of PP without the computational complexity.

This paper describes an interactive visual analytic application that helps one analyze projections in which possibly non-convex densities are well-separated as we focus on looking inside our solution to the multi-class classification problem [11]. In addition, our application identifies axis-parallel projections that are most-closely related to the optimal ones discovered by the algorithm. This is an important tool for analysts as it allows them to describe their results in terms of the original variables rather than uninterpretable linear or non-linear composites of those variables. This aligns with the spirit of visualizing a data mining model to help understand what has been discovered in its context and assess the model's trustworthiness—understanding the probability the model's predictions match new target data and exploring the model's limitations [10].

*E-mail: aanand2@uic.edu
†E-mail: leland.wilkinson@systat.com
‡E-mail: tdang@cs.uic.edu

2 RELATED WORK

PP has been used to find class-separating dimensions [7] and combined with tours to investigate Support Vector Machines [3]. However, these efforts have failed to ameliorate the computational complexity of PP and thus can handle only small datasets.

Using axis-parallel projections of data to find low-D views with large class-separation [9] will be ineffective when there does not exist class separability in axis-parallel views of the data.There are two serious problems with axis-parallel approaches: 1) computations on pairs of dimensions grow quadratically with the number of dimensions, and 2) axis-parallel projections do a poor job of capturing high-dimensional local geometry. While the pursuit of axis-parallel projections can be interesting and may well work for many smaller data sets, there is no reason to expect that they can reveal interesting structures in data. Our platform allows us to test this possibility by displaying optimal projections and their nearest axis-parallel projections to see if these same features can be found in the latter.

3 THE RP EXPLORER

The Random Projection (RP) Explorer has a Projection Viewer on the left and a Variable Viewer on the right as seen in Figure 1. The Projection Viewer shows a horizontally scrollable ranked list of the best 2D random projections that maximize a *Score*. We scale categorical variables as in [11] and generate a set of random 1D projections using three-valued weight vectors with elements $u_j \in \{-1,0,1\}, j = 1,\ldots,p$ for a dataset with p variables. With the classifier [11] example, the Score is a combination of Separation and Purity statistics. For a 1D random projection, our *Separation* statistic, S, is the Euclidean distance between the current-class projected mean $\bar{x}_c$ from the closest other-class projected mean $\bar{x}_k$:

$$S = \min_{k \neq c}(d_{\bar{x}_c,\bar{x}_k})$$

for $k = 1,\ldots,g$ where g is the number of classes in the dataset. We pair the best 1D projections and rank the resulting 2D pairs on a *Purity* measure that favors plots with a large number of pure bins of current class instances for visualization in the RP Explorer.

Each of these ranked plots is a one-against-all classification view with the current class under consideration shown as yellow dots, and the rest of the data shown as gray dots. The scatterplots are binned displays to handle large data [11]. Pure yellow dots represent bins containing only the particular class instances. Pure gray dots represent bins containing only other-class instances. Gray dots with yellow centers and yellow dots with gray centers represent mixed-class bins. There are as many rows as there are classes in the dataset.

Selecting a plot in the Projection Viewer causes the Variable Viewer to be updated to display the two variables "closest" to the projections. We use the *cosine similarity* given by

$$CosineSimilarity = cos(\theta) = \frac{A.B}{||A||||B||}$$

to measure the similarity between the data on an individual variable, seen as vector A, and the random projection of the data, seen as vector B. The p variables are scanned to find the one that maximizes

the similarity to the x-projection on the selected plot and similarly so for the y-projection. This results in the selection of two *exemplar* variables that are shown in the Variable Viewer. Histograms showing the distribution of the variables are oriented based on the dimension they govern—the histogram with bars standing on the y-axis corresponds to the y-dimension variable and similarly so for the other x-dimension. The scatterplot displays the 2D projection of the two data variables and gives the user insight into the separability of the classes in terms of the raw variables.

In some cases, we can see the power of random projections in producing a unique low-dimensional projection that separates classes when even the nearest axis-parallel projection fails to show much separation. In the top plot in Figure 1 showing the Optdigits [2] dataset of optical recognition features of handwritten digits, we can see an instance where the projections reveal better class separation than possible by looking at the most promising variables individually.

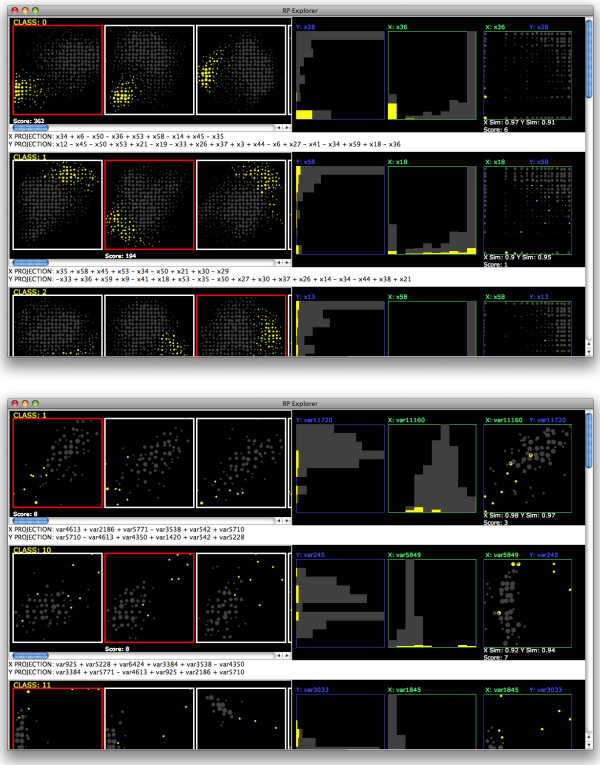

Figure 1: *Top:* RP Explorer displaying class separating projections and variables for the Optdigits dataset. The axis-parallel scatterplots in the Variable Viewer (right panel) do not separate groups as well as the projections in the Projection Viewer (left panel) do. *Bottom:* RP Explorer displaying near-optimal class-separating axis-parallel projections in the Variable Viewer for the Cancer dataset. The random projections in the Projection Viewer are selected to maximize the Classification Score so generally show data clearly separated.

The bottom plot in Figure 1 shows a high-dimensional analysis not amenable to existing visual analytic platforms. There are 16,063 variables in this dataset [8], comprising genetic markers for different types of cancer. With large datasets, those with a large number p of variables, searching for the 2D pair of variables that maximize class separation is a computationally expensive procedure in the order of $O(p^2)$. RP Explorer enables us to identify the most promising markers in this dataset.

Lastly, we compared the automatically discovered important variables discovered by the RP visualizer with those found using manual interaction in [5] for the main example on the E. coli dataset of protein localization sites. Given the exhaustive search done in [5] and the clear separation achieved by the axis-parallel projections, it is validating that RP Explorer came up with the same discriminants.

4 Discussion & Conclusion

The classification problem is not the only potential application of our visual analytic explorer. Our viewer could be programmed to find and show other types of interesting patterns such as unusual densities or features like *Scagnostics* [12]. Our idea is to use random projections to identify highly clustered areas using the Clumpy scagnostic measure. This would produce clustered views of the data similar to the density ranked plots in [9] but would increase our confidence that the axis-parallel projection displayed is closest to the optimal multi-dimensional view. Similarly, we could use the Outlying measure to find axis-parallel projections that have large outliers that could correspond to anomalous events pointing to earthquake scenarios or other climactic events on such climate data.

RP Explorer was inspired by PP, but we must emphasize that it is a computational approximation and does not use the Friedman-Tukey algorithm. Further research is being conducted to determine if PP and our random projection algorithm produce similar results for data involving loss functions not related to classification.

Acknowledgements

This research is supported by NSF/DHS grant DMS-FODAVA-0808860.

References

[1] D. Achlioptas. Database-friendly random projections. In *Proc. of SIGMOD Symposium on Principles of database systems*, pages 274–281, New York, 2001. ACM.

[2] E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. `http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits`, 1998.

[3] D. Caragea, D. Cook, V., and Honavar. Visual methods for examining support vector machine results, with applications to gene expression data analysis. Technical report, Iowa State University, 2005.

[4] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23:881–890, 1974.

[5] A. Hinneburg, D. Keim, and M. Wawryniuk. Using projections to visually cluster high-dimensional data. *Computing in Science Engineering*, 5(2):14 – 25, 2003.

[6] W. B. Johnson and J. Lindenstrauss. Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[7] E.-K. Lee, D. Cook, S. Klinke, and T. Lumley. Projection pursuit for exploratory supervised classification. *Journal of Computational and Graphical Statistics*, 14:831–846, 2005.

[8] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub. Multiclass cancer diagnosis using tumor gene expression signature. *PNAS*, 98:15149–15154, 2001.

[9] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[10] K. Thearling, B. Becker, D. DeCoste, W. D. Mawby, M. Pilote, and D. Sommerfield. *Visualizing data mining models*, pages 205–222. 2002.

[11] L. Wilkinson, A. Anand, and T. Dang. Chirp: A new classier based on composite hypercubes on iterated random projections. In *Proc. of ACM Conf. on Knowledge Discovery and Data mining*, 2011.

[12] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006.