GA-Facilitated Classifier Optimization with Varying Similarity Measures

Michael R. Peterson
Dept. of Computer Science & Engineering
Wright State University
3640 Colonel Glenn Hwy
Dayton, OH 45435
peterson.7@wright.edu

Travis E. Doom
Dept. of Computer Science & Engineering
Wright State University
3640 Colonel Glenn Hwy.
Dayton, OH 45435
travis.doom@wright.edu

Michael L. Raymer
Dept. of Computer Science & Engineering
Wright State University
3640 Colonel Glenn Hwy.
Dayton, OH 45435
mraymer@cs.wright.edu

ABSTRACT

Genetic algorithms are powerful tools for k-nearest neighbors classification. Traditional $k$nn classifiers employ Euclidian distance to assess neighbor similarity, though other measures may also be used. GAs can search for optimal linear weights of features to improve $k$nn performance using both Euclidian distance and cosine similarity. GAs also optimize additive feature offsets in search of an optimal point of reference for assessing angular similarity using the cosine measure. This poster explores weight and offset optimization for $k$nn with varying similarity measures, including Euclidian distance (weights only), cosine similarity, and Pearson correlation. The use of offset optimization here represents a novel technique for enhancing Pearson/$k$nn classification performance. Experiments compare optimized and non-optimized classifiers using public domain datasets. While unoptimized Euclidian $k$nn often outperforms its cosine and Pearson counterparts, optimized Pearson and cosine $k$nn classifiers show equal or improved accuracy compared to weight-optimized Euclidian $k$nn.

Categories and Subject Descriptors

I.5.2 [**Computing Methodologies**]: Pattern Recognition-Design Methodology[classifier design and evaluation, feature evaluation and selection]

General Terms

Algorithms

Keywords

genetic algorithms, pattern recognition, k-nearest neighbors, dimensionality reduction

GECCO'05, June 25–29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

1. INTRODUCTION

The accuracy of some types of classification rules, such as k-nearest neighbors employing Euclidian distance, improves by multiplying the value of each feature by a value proportional to its usefulness in classification [4]. As a method of feature extraction, the application of weights to features in proportion to their classification saliency improves $k$nn classifier accuracy and aids in the analysis of large datasets by isolating combinations of salient features. Through use of a bit-masking feature vector, GAs have successfully performed feature selection in combination with a $k$nn classifier [5].

More recently, cosine similarity has successfully been employed as an alternative similarity measure to euclidian distance for $k$nn classification. Adjustment of feature weights may improve classifiers employing cosine similarity [2], which assesses the angular closeness of two feature vectors, taken relative to a point of reference (i.e. the origin). Changing the point of reference changes the similarity between vectors also affects the performance of a cosine-based $k$nn classifier. Peterson et. al. [3] employ a GA to simultaneously optimize feature weights and the point of reference (i.e. feature offsets) for cosine $k$nn. The GA searches for an optimal point of reference to assess the angular similarity. They report classification results highly competitive with contemporary classification techniques including support vector machines, feed-forward neural networks, and decision tree algorithms.

The authors present a novel form of classifier optimization for $k$nn classifiers employing Pearson correlation as a similarity measure. Like cosine similarity, Pearson correlation is frequently used as a similarity measure to classify cancer tissues using gene expression data [1]. Careful selection of feature weights in proportion to feature saliency improves the performance of Pearson $k$nn. Additionally, applying offsets to features affects classification. Applying an additive or subtractive shift of a single measurement (i.e. feature) affects the correlation between two feature vectors. Allowing a GA to search for an optimal set of offsets allows maximization of within-class pattern correlation and minimization of between-class correlation, hence improving Pearson correlation-based $k$nn accuracy.

2. METHODS

K-nearest neighbors classifiers employing Euclidian distance, cosine similarity, and Pearson correlation are opti-

mized by a GA for class accuracy, the balance between accuracies of each class, and the number of features used. For each classifier, the GA searches for an optimal set of feature weights and a k-value. Weights on the chromosome range from 0.0 up to 100.0, and are normalized by sum to 1 and applied to feature values before classification. Feature values in datasets have been normalized to range between 1.0 and 10.0 to avoid bias between features. The k-value ranges from 0 to 30, 50, or 100, depending on the size of the dataset. For Pearson and cosine $k$nn, offsets for each feature are also evolved. Offsets are allowed to shift feature values in the range -15.0 and 25.0 before feature weights are applied. Offsets are not applied to Euclidian $k$nn because Euclidian distance is invariant to feature shifting.

The Pearson correlation between feature vectors $\vec{x}$ and $\vec{y}$ is

$$pear(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{f}(x_i - \bar{x})(y_i - \bar{y})}{(f-1)S_x S_y}$$

where f is the number of features, $\bar{z}$ is the mean value of vector $\vec{z}$, and S_z is the standard deviation of $\vec{z}$. The pearson correlation coefficient measures the strength of a linear relationship between $\vec{x}$ and $\vec{y}$, and ranges from -1, indicating a strongly inverse linear relationship, up to +1, indicating a strongly positive linear relationship. A coefficient value of 0 indicates the absence of any detectable relationship between vectors. The experiments described here only consider positively correlated training patterns similar. Equations for Euclidian distance and cosine similarity, as well as the specific technique for assigning class labels, is described in [3].

The experiments employ the GA and fitness function described in [3]. 20 optimizations each are performed for Euclidian $k$nn weights, cosine $k$nn weights, cosine $k$nn weights + offsets, Pearson $k$nn weights, and Pearson $k$nn weights + offsets. Datasets are split randomly into training, test, and bootstrap validation sets for each replication. For comparison purposes, 20 replications using unoptimized Euclidian, cosine, and Pearson $k$nn are also performed.

3. RESULTS

Figure 1 shows boxplots of accuracies obtained for each classifier for the Pima diabetes dataset obtained from the UCI machine learning repository. Though cosine and Pearson $k$nn perform worse than Euclidian $k$nn without any training applied, their performance improves when applying weight-only optimization. Further improvement for both algorithms occurs when simultaneously optimizing both weights and offsets, to a level competitive with weight-optimized Euclidian $k$nn. Performance surpassing that of weight-optimized Euclidian $k$nn has been obtained for other public domain datasets. Though not shown here, GA optimization also reduces bias towards larger classes by placing selective pressure on balanced classification and reduces the dimensionality of the data, removing features of low or no relevance.

4. CONCLUSIONS

In conjunction with weight optimization, offset optimization represents an effective method for improving the performance of k-nearest neighbors classifiers employing cosine similarity or pearson correlation. For some datasets, offset optimization improves classification accuracy and balance over weight-only optimized classifiers. While unoptimized

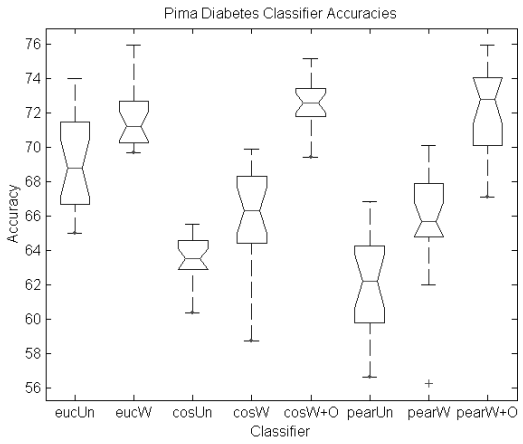

Figure 1: Boxplots comparing classifier accuracy distributions for Pima diabetes.

euclidian $k$nn classification often outperforms unoptimized cosine or person $k$nn classification, the GA-facilitated optimization techniques presented in this paper allow the cosine and pearson $k$nn classifiers to match or outperform optimized euclidian $k$nn classifiers. The weight optimization performed by the GA maintains feature independence while discovering relative feature importance, thus potentially providing novel insight into the problem domain under consideration.

5. REFERENCES

[1] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. A. Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Science*, 97:262–267, 2000.

[2] E. Han, G. Karypis, and V. Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In *Advances in Knowledge Discovery and Data Mining: fifth Pacific-Asia Conference*, pages 53–65, 2001.

[3] M. R. Peterson, T. E. Doom, and M. L. Raymer. Ga-facilitated knowledge discovery and pattern recognition optimization applied to the biochemistry of protein solvation. In *GECCO 2004 Proceedings, LNCS 3102*, pages 426–437, 2004.

[4] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain. Dimensionality reduction using genetic algorithms. *IEEE Trans Evol. Comp.*, 4(5):164–171, 2000.

[5] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pat. Rec. Letters*, 10:335–347, 1989.