CHIRP: A New Classifier based on Random Projections and Covering

Anushka Anand *aanand2@lac.uic.edu*
Leland Wilkinson *Leland.wilkinson@systat.com*
Dang Nhon Tuan *tdang@cs.uic.edu*
Department of Computer Science, College of Engineering, University of Illinois at Chicago

UIC Department of Computer Science
UNIVERSITY OF ILLINOIS AT CHICAGO
COLLEGE OF ENGINEERING

Introduction

We introduce a classifier called CHIRP based on the L∞ norm. This algorithm is an iterated sequence of projecting, binning and covering designed to overcome the curse of dimensionality, computational complexity and nonlinear separability. It outperforms leading competitors in prediction accuracy on widely-used benchmark datasets.

The Problem

<u>*Supervised Classification*</u>: Given a set of training data pairs,

$T = \{(x_1,y_1),\ldots,(x_n,y_n)\}$

with *feature vectors* $x_i = [x_{i1},\ldots,x_{ip}]$ where p = # variables
and *class labels* $y_i \in \{y_1,\ldots,y_g\}$ where g = # classes
Find a classification function f: X → Y that maps x ∈ X to y ∈ Y that maximizes the number of accurate predictions of y for unseen input feature vectors x.

Hypercube Description Regions

A *Hypercube Description Region* (HDR) describes the set of points less than a fixed distance from a single center point using the *L∞ norm,* defined for a vector x as:

$\|x\|_\infty = \sup(|x_1|,\ldots,|x_n|)$

Generally, the weighted formulation for finite dimensional vectors is

$\|x\|_\infty = \max(w_1|x_1|,\ldots,w_n|x_n|)$

Composite Hypercube Description Regions (CHDRs) are unions of HDRs that define large-scale structures as shown in Figure 1.

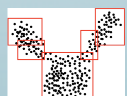


Figure 1: CHDR covering a non-convex geometric object. [from B. J. Gao and M. Ester, "Turning clusters into patterns: Rectangle-based discriminative data description," in ICDM 2006]

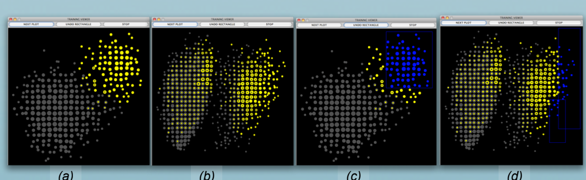
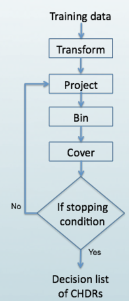
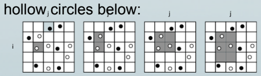


Figure 2: Examples of CHDRs covering pure bins in the CHIRP algorithm. (a) and (b) show 2D views of bin matrices of some random projection of the data. Yellow points are the target class, gray points are other classes. Each bin is represented by a dot at the centroid of points falling in the bin and the size of the dot is proportional to the count of points in the bin. Yellow points with gray centers and gray points with yellow centers reveal overlapping densities. (c) and (d) show CHDRs as blue rectangles covering regions of bins.

The CHIRP Algorithm



- *Transform:* Transform numerical variables to reduce extreme skewness.
- *Project:* Pick a target class (go through all classes in a cyclic order). Compute random unit-weighted projections of all variables and rank the projections on a Separation Index – to favor projections with large separation of the target class from all other classes.
- *Bin:* 2D bin the data on the chosen 2D projections to deal with large datasets. Rank the bin-matrices on a Purity measure based on the number of pure bins of the target class.
- *Cover:* Grow a CHDR covering pure bins by expanding in a spiral path. Pure bins are hollow circles below:



- *Stop:* Return the list of CHDRs when no more CHDRs can be grown.
- *Ensemble:* CHIRP is an ensemble classifier – run 7 instances and take the majority vote.
- *Score:* Transform and project a new data point using training data parameters. Classify a point based on the CHDR it is contained in or closest to using the L∞ distance.

Experiment

CHIRP was tested on 20 datasets (from the UCI Machine Learning Repository and Statlog), each posing a different challenge to classifiers due to their structural variety.

CHIRP is included in the Weka Data Mining Workbench and tested against 50 classifiers implemented in Weka.

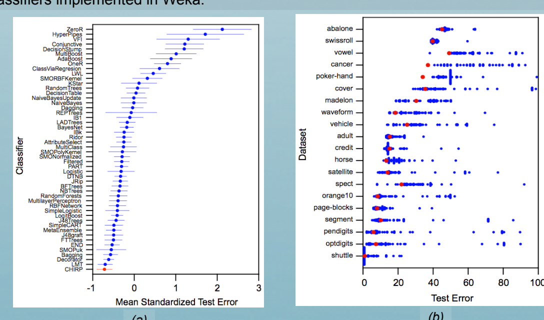


Figure 3: Classification accuracy of CHIRP versus other Weka classifiers. (a) shows the mean standardized test error with 95% confidence intervals within 20 datasets for all classifiers. (b) is a dot plot showing test errors of all classifiers on all 20 datasets with CHIRP as a red dot.

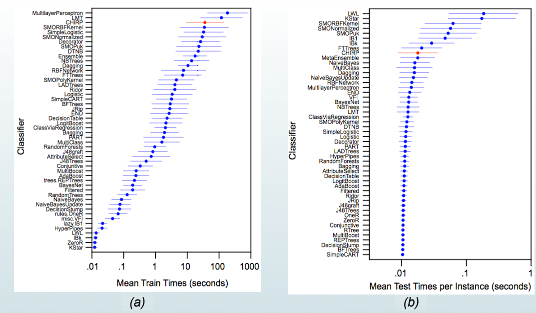


Figure 4: Efficiency of CHIRP versus other classifiers showing mean times with associated 95% confidence intervals within the 20 datasets for (a) training and (b) testing.

Discussion

Figure 3 & 4 show CHIRP is a relatively efficient classifier that is unusually accurate with consistent performance over a variety of different datasets. Intuitions of why CHIRP works so well are:

- CHIRP embeds projecting, covering *inside* iterations – it peels already classified cases at each iteration of the training to reveal hidden regions of density.
- CHIRP handles non-convex, discrete and disjoint densities by covering.
- CHIRP's affine 2D projections provide non-orthogonal views of densities and deal with curse of dimensionality.
- The two fitness measures (Separation & Purity) target different aspects of densities. Separation values projections with large margins and Purity values projections with compact subsets so good projections missed by one will likely be found by the other.
- CHIRP's unit weights have good out-of-sample performance – a form of regularization.
- It is an ensemble classifier (like Random Forests) - *boosts* performance

Conclusions

Highlights of the CHIRP algorithm:

- CHIRP has the lowest average standardized testing error rate and achieved the lowest error rate on more datasets than did any other classifier.
- There is no need to dummy-code variables; categorical variables expend 1 degree-of-freedom.
- CHIRP is linear in complexity on *n* (number of cases) and *p* (number of variables) an d subquadratic on *g* (number of classes).
- CHIRP does not depend on sensitive adjustable parameters.
- CHIRP is readily parallelizable.
- CHIRP is a novel algorithm - not a hybrid classifier - therefore uniquely suited for applications with limited a priori knowledge of the data and for inclusion in ensembles.