

Title

Hello my dear professors,

I'll be presenting my master's thesis, titled Spectral methods for outlier detection in machine learning

First, I'd like to present a short overview of our work.

Next Slide - Overview

Our work deals with the problem of outlier detection

We argue that spectral methods are valuable

Propose to combine spectral and outlier detection methods

Evaluate our approach on 20 data sets and discuss the results

Give a short outline;

First;

we analyze the problem of outlier detection

And outlier detection methods

Then

we move on to spectral methods and review some of them

After we present our idea,

we move on to experiments where we evaluate the performance of outlier detection methods, and their combination with spectral methods

So how can we define outliers?

Next slide

What is an outlier?

A popular definition by Grubbs:

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs

For example, here this point is quite different from others

however note that it is not easy to define what makes an instance outlier

are these outliers?
Why important?

Convey valuable and actionable information in real life
Ex: a network attack, potential fault, disease, cancerous tumor

From the perspective of learning theory, makes it possible to learn better models.

Now we look how outlier detection differs from classification

Next Slide - How it differs from classification?

availability of labels
supervised, semi-supervised, unsupervised

class priors are unbalanced
classification cost unsymmetric
noise is similar to outliers

Now we look at the different kinds of methods for outlier detection

Next Slide - Classification of Methods

Learns a discriminative model that separates outlier and typical instances
Requires labeled data
Two class vs One class

Density Estimation
Assumes outliers occur far from typical instances, in low density regions

High computational complexity and low performance on high-dimensional inputs
Parametric, Semi-Parametric, Non-Parametric

Statistical Methods, Nearest neighbor methods, Clustering methods

Now we review the outlier detection methods we use in our work

Next Slide - Active Outlier

Supervised, One class method

Reduces unsupervised outlier detection problem to classification of normal samples from artificially generated outliers

An ensemble of classifiers are trained on selectively sampled subsets of training

Requires much less computational power when compared to density estimation and spectral methods

Next Slide – Active Outlier I

this is our training data, only typical

we want to learn a tight boundary around it

Next Slide – Active Outlier II

we draw instances from a background dist. And assume they are from a diff. class

Next Slide – Active Outlier III

next we train a classifier, here a decision tree

now we find the regions where the classifier makes most mistakes and re sample from those regions

Next Slide – Active Outlier IV

and train another decision tree until we are satisfied

now we move on to the other outlier detection method

Next Slide - Local Outlier Factor

Nearest neighbor based

Considers differences in local densities around an instance

we calculate a local reachability distance for each instance, measuring the difference between k-distances of its neighbors

if k-distances differ much, lrd is low

for a point, LOF is the sum of the ratio of lrd of neighbors to its lrd value.

If lrd of neighbors are high and ours is low, we are an outlier

Hard to find an optimal k , depends on problem

LOF values are quite sensitive to k value

Calculate LOF values for k in k_{\min} , k_{\max} and take maximum

next, we review another outlier density based outlier detection method

Next Slide - Parzen Windows

Non-parametric density estimation method

Instead of using hard counts in our histogram estimation, we use soft counts

Determining an optimum bin size is difficult

A fixed value of bin size for the whole input space may not work well

Next Slide – Active Outlier III

Algorithm that returns a function f that takes the value 1 in a “small” region capturing most of the data -1 elsewhere.

strategy is to map the data into the feature space

corresponding to the kernel and to separate them from the origin with maximum margin

ν parameters controls the number of outliers, support vectors

ν is an upper bound of the fraction of outliers

lower bound on the fraction of SVs

We have reviewed outlier detection methods we use in or work now we move onto Spectral Methods

Next Slide - Formal Definition

Unsupervised learning techniques that reveal low dimensional structure from high dimensional data

Use the spectral decomposition of specially constructed

matrices to reduce dimensionality and transform input data to a new space

Linear vs. Nonlinear

Generic approach Find a lower dimensional representation for input data where the dot products in this new space matches the similarities as soon as possible. this is also low rank matrix approx problem. solution is svd=spectral composition

We review 4 spectral methods, we give the cost functions they minimize or maximize

Next Slide – Spectral methods Costs

first, PCA, we want to find a transformation matrix such that reconstruction error minimum=variance is maximum
solution is eigenvectors of covariance matrix

we can apply Kernel Trick to PCA, resulting in Kernel PCA
now we have a nonlinear method, we directly apply spectral decomposition on kernel matrix to obtain transformed points Z

like PCA, Kernel PCA requires input to be centered, zero mean
we can center in the kernel space with this operation

LEM,

Map similar instances to closer points in new space

Assumes that each input instance denotes a node in a graph and uses the adjacency matrix as a similarity matrix

Each instance is connected to only a small number of close instances resulting in a sparse adjacency matrix

Cost function gives more weight to close points, we take the smallest eigenvectors living the 0 eigenvalued one

MDS

A family of dimensionality reduction methods, we discuss the Classical (or Metric) MDS

The aim is to preserve the Euclidean distances in the projected space as much as possible

we extract dot products from euclidean distance matrix and use the spectral decomposition to get Z

Now we look at a visual example and point out a few properties of these methods

Next Slide – Visual Example

$k=2$, gaussian

PCA does not change anything,

LEM maps to close points

KPCA, kernel matrix positive, so all instances are in same half space

KPCA – M_c and MDS are equivalent, mean subtraction transforms the problem from dot products to euclidean distances

we finished reviewing spectral methods, now we present our idea

Next Slide – Spectral Outlier Detection, Idea

we propose combining spectral methods with outlier detection, LEM and MDS

first apply spectral method then apply outlier detection

why?

Curse of dimensionality

- Distance functions lose their meaning in high dimensions
- Processing a higher dimensional data requires more computation

New attributes come from combinations of the original ones, methods can use multivariate information.

Spectral methods transforms possibly complicated patterns to smoother ones, discriminative methods fit simpler boundaries

Next Slide – Revealing low dim structure

here we see that nonlinear pattern which requires a complex boundary, is transformed to a linear one, where we learn a much simpler boundary

Now we continue with the evaluation of outlier detection methods and our spectral outlier detection algorithm

Next Slide – Evaluated Methods

Slayttan anlat

Before we go on to results on 20 data sets we observe the applicability of our approach on a face detection problem

Next Slide – Face Detection

Slaytı anlat

We first apply PCA and AO on the data set, here are the results

Next Slide – Face PCA

AUC=.76,

see that instances overlap a lot

If we apply LEM first and then AO

Next Slide – Face LEM

AUC increases to .97 and much less overlap

now we present our evaluation results on 20 data set,
first we look at experiment setting

Next Slide – Experiment Setting I

approach as a rare class problem, take classification data sets, and form outlier detection problems with the majority class and other small sized classes

Priors unbalanced, so measure AUC under ROC

we find the best parameter combination on train/val set

and obtain 10 AUC values on test set with CV

we carry out, semi-supervised and unsupervised experiments

For the similarity matrices we use

Next Slide – Experiment Setting II

3 kernels

neighbor count, variance,

d for PCA and LEM, MDS (4 linearly spaced values)

how do we analyze results?

Next Slide – Experiment Setting III

we first look at outlier detection methods individually and then their combination with spectral methods,
we apply Friedmans with nemenyi on accuracies
5X2 CV F on pairs
and sign, Wilcox sign on wins

before we move on to results, we look at the data sets

Next Slide – Data sets I

kolonların ne olduğunu söyle
as you see data sets show a great variation
some have large sample some low
some are very high dimensional, some are low
also in some cases, percent of outliers is very little but in some cases it is too much to maybe consider them outliers

now we compare outlier detection methods
we look at all four methods first with no transformation

Next Slide – No transformation – Unsupervised

x axis, data sets
y axis, AUC
vertical bars, 1 std. deviation
we see performamnces of AO, LOF, PW, SVM

there is no clear winners, some are better at some data sets, some are at others
but we see that AO is generally not good, and it has a high std deviation

Next Slide – No transformation – semi – supervised

again no clear winners but performamnces are much closer

now we look at the results of sign tests

Next Slide – No transformation – WTL

wins of algorithm on row against the alg. in column
if sign test signif, bold
if wilcox sign, *
no sign diffs

Next Slide – PCA – Unsupervised

again no clear winners, some are better at some data sets, some are at others
now LOF is also not good

Next Slide – PCA – semi – supervised

again no clear winners but performances are much closer
now we look at the results of sign tests

Next Slide – PCA– WTL

no significance in semi-supervised
but svm is better than lof for unsupervised with wilcox

Next Slide – LEM – Unsupervised

SVM is always top
but LOF and PW are close
AO worst

Next Slide – LEM – semi – supervised

again no clear winners but performances are much closer

Next Slide – LEM– WTL

no significance in semi-supervised
but svm is better than ao for unsupervised with wilcox and sign

Next Slide – MDS – Unsupervised

similar to LEM

MDS and AO do not go well together

Next Slide – MDS – semi – supervised

again no clear winners but performances are much closer
AO still not good

Next Slide – LEM– WTL

in semi-supervised, $svm > ao$ for sign and wilcox
for unsupervised, no sign but nearly $svm > ao$
now we summarize these results

Next Slide – Significant Diff between Outlier Detection methods

here are the nemenyi results,
methods on left are better, with smaller average rank
if connected no sign difference

not one best algorithm but SVM is always best
MDS+AO peculiar, always very bad results

now we compare spectral methods to see which spectral method is better
for outlier detection

Next Slide – AO Unsupervised

a similar situation, heavily data set dependent

Next Slide – AO Semisupervised

performances are closer but no clear winners

Next Slide – AO WTL

no significance, lots of ties

Next Slide – LOF Unsupervised

LEM and MDS either better or equal in performance

Next Slide – LOF Semisupervised

except PCA all are close

Next Slide – LOF WTL

no signif in semi-supervised
but lem>pca, lem>orig and mds > pca

Next Slide – PW Unsupervised

data set dependent

Next Slide – PW Semisupervised

closer

Next Slide – PW WTL

no signif in both cases

Next Slide – SVM Unsupervised

less data set dependent, all are closer but lem or mds are always good

Next Slide – SVM Semisupervised

closer

Next Slide – SVM WTL

although no significance, SVM-lem and SVM-MDS in unsup. Has high wins

Next Slide – Signif Diff between Spectral Methods

for semi-supervised, there are no diff, because all methods are able to learn when given only typical

in unsupervised, for AO and PW, trans. Is not important.

We may find some parameters config for each that always give best perf.

AO learns decision tree, PW nn,

however, for LOF and SVM, LEM and/or MDS gives significantly better results

now we wrap up these results and conclude

Next Slide – Conclusions

we see that there is no best outlier detection method, perf depends on data set

but SVM is good, it performs comparably and has low comput.

Complexity

especially semi-supervised case, everyone can learn a good model

but unsupervised scenario is more realistic as we do not have labels most of the time

for unsupervised, AO and PW learn no matter what the transformation is, but for SVM and LOF we see the contribution of spectral methods, they enable these methods to perform better,

note that SVM and LOF are good methods compared to AO and PW, so SVM+LEM is a good algorithm

in real life choosing an algorithm depends on data set,

do we have labels, is it high dimensional, online or offline running time etc.

Next Slide – Future Work

we have taken a random data set collection

we think that SVM and LOF with LEM may be good, we need to analyze on which data sets their perf is good/bad

are they both good/bad on same data sets?

Then we can say that spectral methods are valuable in these data sets and show it by finding more data sets

other future work, try ISOMAP or LLE

For PCA we use only typical to find transformation but

LEM and MDS function unsupervised manner, do not use labels

find the transformation from only typical sample

modify them to use only typical

interpolate new instances using transformed typical instances', find nearest take average etc.

then there may be signif differences in semi-supervised case too.