A comparison of three class separability measures

L.S Mthembu & J.Greene

Department of Electrical Engineering, University of Cape Town
Rondebosch, 7001, South Africa.

lsmthlin007@mail.uct.ac.za & jrgreene@eng.uct.ac.za

Abstract

Measures of class separability can provide valuable insights into data, and suggest promising classification algorithms and approaches in data mining. We compare three simple class separability measures used in supervised machine learning.

Their relative effectiveness is evaluated through their functional relationships and their random projections of data onto R^2 for visualization.

We conclude that the simple direct class separability measure of a dataset is an easier and more informative measure for separability than the class scatter matrices approach and it correlates well with Thornton's Separability's index.

1. Introduction

In exploratory analysis of data, simple and rapidly computable global measures such as class separability can give insights into the data and provide pointers towards the choice of classifier.

Given any dataset, one would like to know how separable the classes are before choosing a particular classifier. For low dimensional datasets (≤3) we can view the class scatter.

Unfortunately most real world datasets have more than three dimensions – furthermore we would like to automate this procedure by minimizing user-chosen free parameters in all the measures.

We compare the following three data dependent class separability measures:

1) Class Scatter Matrices (CSM)
2) Thornton's Separability index (Sepindex, SI)
3) Direct Class Separability measure (DCSM)

The class scatter matrices [1] approach is a well-known and widely used measure (particularly in the context of clustering). However this measure aggregates cluster separation into a measure based on the separation of means and thus all class distribution information is lost.

In the previous paper [2], Thornton's SI was shown to be an effective measure of class separability, well suited to feature selection in nearest neighbour and kernel classifiers.

It is possible to define a direct measure of separability in which mean distance is replaced by summation of individual pairwise distances.

The present paper examines the hypothesis that such a measure, retaining as it does distribution information may be more informative than the class scatter matrix measure. We call such a measure the direct class separability measure (DCSM).

1) **Class scatter matrices/measure** (**CSM**) for class separability is an old technique. It is defined as:

$$S_b = \sum_{i=1}^{c} (m_i - m)^t (m_i - m)$$

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{ni} (x_{ij} - m_i)^t (x_{ij} - m_i)$$

where: *c= number of classes, ni = number of instances in class i.* m_i *is the mean of instances in class i and* m *is the mean of all classes.* x_{ij} *is the jth instance in class i.* S_b *is the between class scatter matrix and* S_w *is the within class scatter matrix.*

$$J = \text{trace}(S_B) / \text{trace}(S_w)$$

J is an unbounded measure. The larger the value of *J* the smaller the within class scatter as compared to the between class scatter.

2) **Separability Index** (**SI**) as defined in [2]:

$$\mathrm{SI} = \frac{\sum_{i=1}^{n} (f(x_i) + f(x_i') + 1) \bmod 2}{n}$$

calculates the average number of instances that share the same class label as their nearest neighbours. The performance of Thornton's separability index has been previously demonstrated in [2]. We thus report on the functional relationship of the other two separability measures versus this index.

3) We define the **Direct class separability measure, DCSM** to be:

$$S_w = \Sigma_{i=1}^{ni} \; \Sigma_{i=1}^{ni} \left\| x_i - x_j \right\|$$

$$S_b = \Sigma_{i=1}^{ni} \; \Sigma_{j=1}^{nj} \left\| x_i - x_j \right\|$$

where ni & nj = number of instances in class i & j respectively and x_i & x_j are the instances.

$$DCSM = [S_{\boldsymbol{B}} - S_w].$$

$S_{\boldsymbol{B}}$ is the between class distances
S_w is the within class distances

This measure directly measures how compact each class is as compared to how far it is from the other class.

If for one dataset, $S_B < S_{w-}$ and $S_B > S_{w+}$ then the scatter of the negative class is more than the scatter between it and the positive class. Further more, the negative class overlaps the positive class.

One way of comparing correlation between separability measures is via feature selection. This is presented in section 2 of this paper.

To further explore the differences between these measures section 3 presents the all measures' random projections of a number of datasets onto two-dimensional space. Section 4 presents conclusions of the paper.

2. Functional Relationships

We calculate each measure on all 2^d-1 feature subsets of each dataset, where d is the number of features in a dataset. We then plot the value of each separability measure versus the value of Thornton's separability index.

We make use of the Wisconsin Breast -Cancer and Liver datasets from [3], the Ljubljana Breast-Cancer and Thyroid from [4]. We arbitrarily used realization 13 on the datasets from [5].

Comments on functional relationship graphs

The class scatter matrices (CSM) vs. SI figures show that CSM does not have a clear functional relationship with Thornton's separability index.

When the class scatter matrices measure has a feature set that produces the best class separability, the SI does not. This is different when we compare the plots of the direct class separability measure (DCSM) graphs.

It is found that there is a clearer correlation between the direct class separability measure and Thornton's separability index than there is with the class scatter matrices; furthermore one of the classes in the DCS measure can have an inverse (negative slope) relation with SI. This is additional information given by using this measure.

The definition of DCSM means we will generate two graphs of this measure for each dataset. The first graph, DCSM vs. SI S_w+, for example, shows how the measure varies, for the within class scatter distances (S_w+) of the positive class for different feature combinations.

When the slopes of the relationship between DCSM and SI for the positive and negative classes are the same, the classes are easily separable. This separability *can* be in the form of multi-clusters within each class (multimodal) and or uni-modal (each class being one compact cluster).

This results from the fact that the distances between the positive class instances are smaller than the distances between the positive class instances are from the negative class instances and the distances between the negative class instances are smaller than the distances between the negative class instances are from the positive.

When the slopes of the relationship between DCSM and SI for the positive and negative classes are different, one of the classes is overlapping the other. This is due to the fact that the one class has within class distances that are larger than its instances are from the opposing class's instances.
The class scatter matrices approach does not explicitly tell us this information.

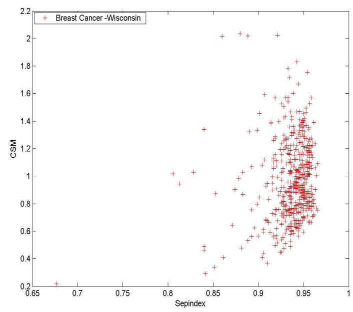

Figure 2.1 CSM vs. SI
(*B-Cancer Wisconsin*)

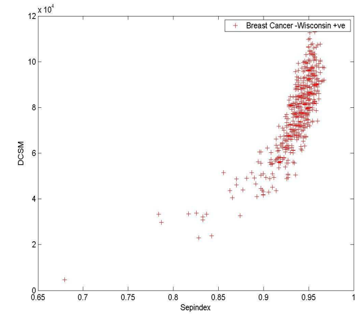

Figure 2.2. DCSM vs. SI (S_w+)
(*B-Cancer Wisconsin*)

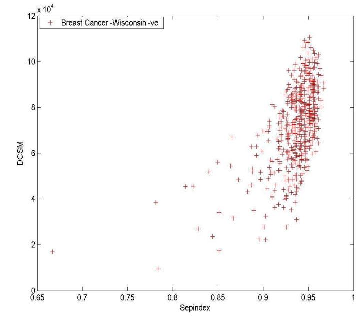

Figure 2.3 DCSM vs. SI (S_w-)
(*B-Cancer Wisconsin*)

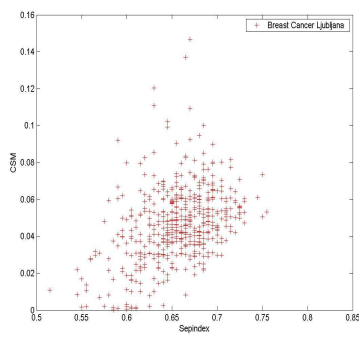

Figure 2. 4 CSM vs. SI
(*B-Cancer Ljubljana*)

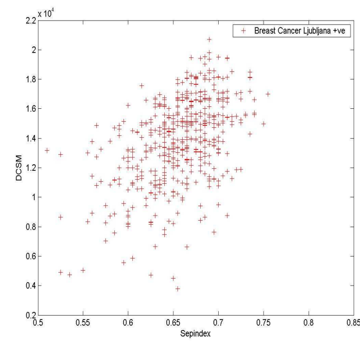

Figure 2.5. DCSM vs. SI (S_w+)
(*B-Cancer -Ljubljana*)

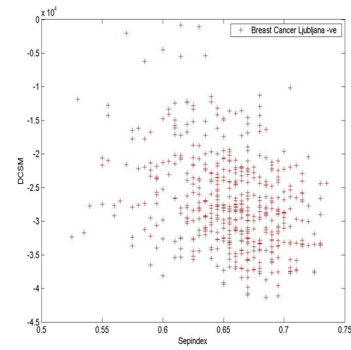

Figure 2.6. DCSM vs. SI (S_w-)
(*B-Cancer Ljubljana*)

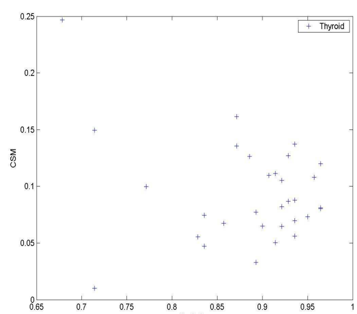

Figure 2.7. CSM vs. SI
(*Thyroid*)

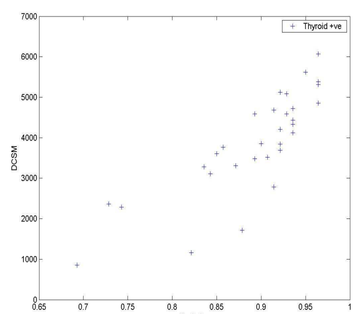

Figure 2.8. DCSM vs. SI (S_w+)
(*Thyroid*)

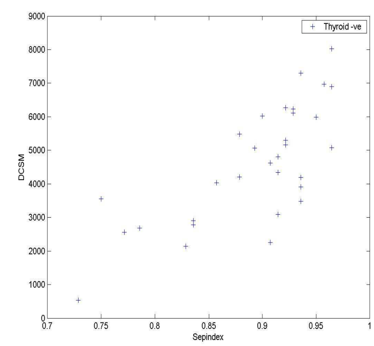

Figure 2.9 DCSM vs. SI (S_w-)
(*Thyroid*)

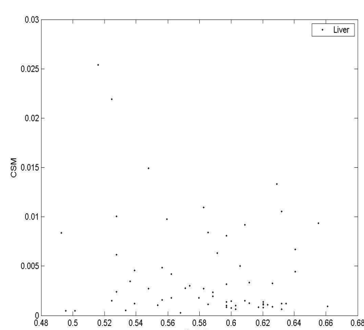

Figure 2.10. CSM vs.SI
(*Liver*)

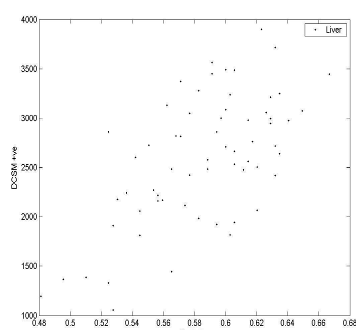

Figure 2.11. DCSM vs. SI (S_w+)
(*Liver*)

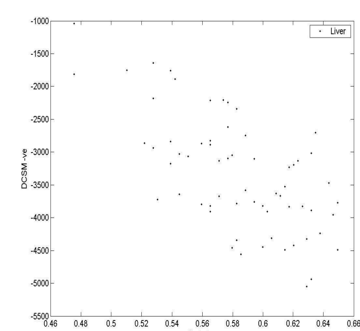

Figure 2.12.DCSM vs. SI (S_w-)
(*Liver*)

3. 2D Projections of Datasets using the above measures to maximize separability

We project the full dataset onto a 2 dimensional space by 2 random vectors: Given an X = [*pxd*] observation matrix, we multiply it by a random matrix R= [*dx2*]. This projects the *d* dimensional data onto 2 dimensional space.

We generate 100 random (dx2) vectors and plot the graph that maximizes the separability measure in question.

Comments on random projection figures

The projection graphs confirm the results from the functional relationship graphs of the previous section.

When the classes in the dataset are distinct from each other the DCSM vs. SI functional relationship slopes remain the same for both classes (e.g. Wisconsin B-Cancer and Thyroid). When the slopes differ the classes in the dataset are not easily separable due to classes overlapping (e.g. Liver and Ljubljana B-Cancer).

This measure in effect tells us before hand how our classes are possibly distributed in relation to one another.

Thornton's separability index also tells us about how much class overlap there is; the more overlapping between the classes the more instances will have nearest neighbours of a different class, resulting in a low SI.

Interestingly both the above measures do not explicitly tell us whether the classes are multimodal or just uni-modal but only tell us the degree of overlap in the classes.

It is thus not surprising then that the direct class separability measure's projections are similar to Thornton's Separability index's projections of each dataset

Figures 3.8 to 3.9 show the multimodality of the Thyroid data while figure 3.7 does not clearly show this structure in the data. This is because the CSM aggregates the instances and their classes thus the information of the diversity of the data structure is lost.
All three separability measures are not able to produce projections of separable classes for the Liver and the Ljubljana breast cancer datasets. This was alluded to by the low SI index and the change in slope on the DCSM vs. SI graphs in the previous section, meaning the above mentioned datasets are not easily separable.

4. Conclusions

We have compared three class separability measures used in machine learning; class scatter matrices (CSM), direct class separability (DCSM) and Thornton's separability measure (SI).

We have shown that the CSM measure does not have a clear functional relationship with the SI while the direct class separability measure does.

The lack of good correlation between CSM and SI is due to the loss of structural information (due to the averaging of instances and classes) in the evaluation of the class scatter matrices measure. This measure is biased to Gaussian compactly clustered classes. It does not work well with multi-clustered classes.

DCSM on the other hand gives further information on the structure of the classes, i.e. their compactness and whether one class overlaps the other or not, by the inverse or direct relationship with the SI measure.

The more separable the classes are, the more direct (i.e. $S_B > S_w+$ & $S_B > S_w-$) the relationship between DCSM and SI as opposed to the inverse relationship (i.e. $S_B > S_w+$ or S_w-) for a non-easily separable dataset.

Direct class separability (DCSM) as opposed to the class scatter matrices (CSM) is a quick and more informative method of extracting information about the class scatter of a dataset.

References

[1] Pattern Classification and Scene Analysis. Richard .O Duda *John Wiley and Sons* 1973.

[2] Feature Subset Selection using Thornton's separability index and its applicability to a number of sparse proximity-based classifiers. J.Greene 2001 *PRASA 2001*.

[3] University of California, Irvine Machine Learning Database Repository at: www.ics.uci.edu/~mlearn/mlrepository /html

[4] http://www.first.fraunhofer.de/~raetsch/ by G.Rätsch.

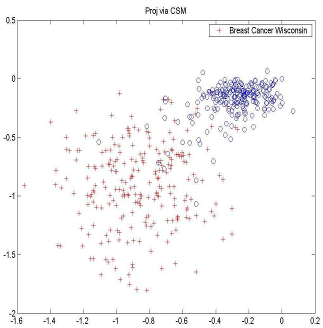

Figure 3.1 Projecting B-Cancer Wisconsin via CSM

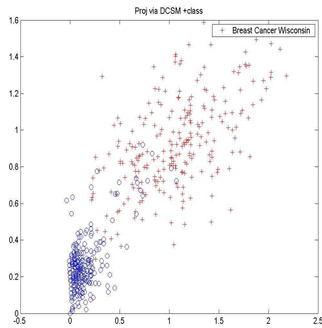

Figure 3.2 Projecting B-Cancer Wisconsin via DCSM

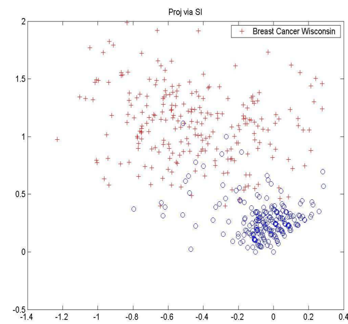

Figure 3.3. Projecting B-Cancer Wisconsin via SI

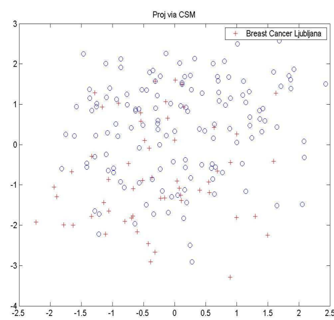

Figure 3.4 Projecting B-Cancer (Ljubljana) via CSM

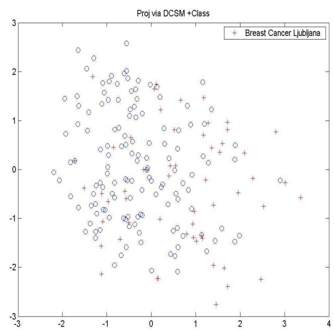

Figure 3.5 Projecting B-Cancer (Ljubljana) via DCSM

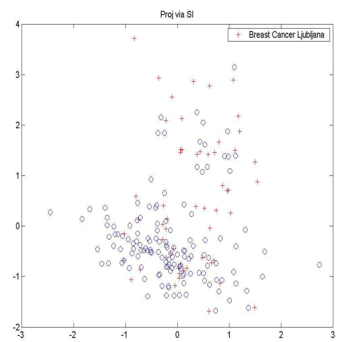

Figure 3.6 Projecting B-Cancer (Ljubljana) via SI

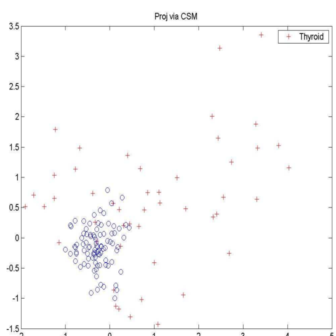

Figure 3.7 Projecting Thyroid via CSM

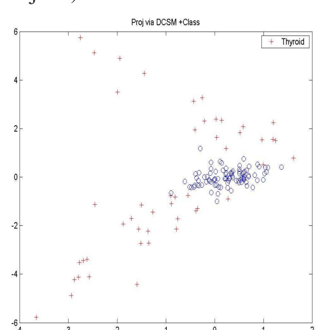

Figure 3.8 Projecting Thyroid via DCSM

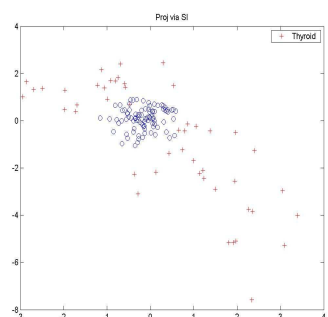

Figure 3.9 Projecting Thyroid via SI

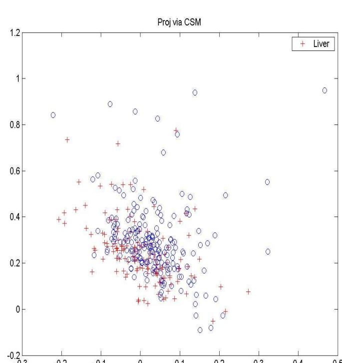

Figure 3.10 Projecting Liver via CSM

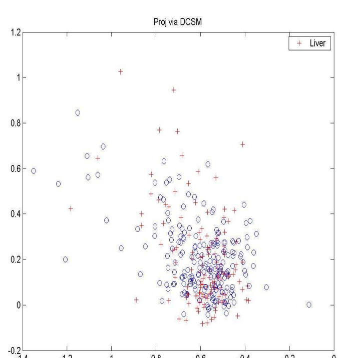

Figure 3.11 Projecting Liver via DCSM

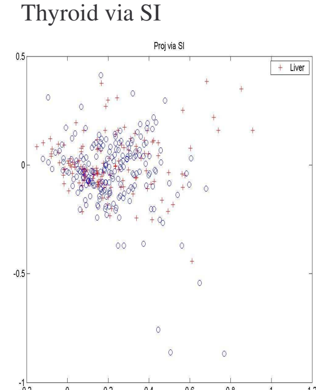

Figure 3.12 Projecting Liver via SI