A New Discernibility Metric and Its Application on Pattern Classification and Feature Evaluation

Zacharias Voulgaris[1]

[1] Independent Contractor, 7750 Roswell Rd. #3D,
Atlanta, Georgia 30350, USA
research@voulgaris.tk

Abstract. A novel evaluation metric is introduced, based on the Discernibility concept. This metric, the Distance-based Index of Discernibility (DID) aims to provide an accurate and fast mapping of the classification performance of a feature or a dataset. DID has been successfully implemented in a program which has been applied to a number of datasets, a few artificial features and a typical benchmark dataset. The results appear to be quite promising, verifying the initial hypothesis.

Keywords: Discernibility, Feature Evaluation, Classification Performance, Dataset Evaluation, Information Content, Classification, Pattern Recognition.

1 Introduction

The potential of accurate classification of a feature or a set of features has been a topic of interest in the field of pattern classification. Especially in cases where the classification process is a time-consuming or generally impractical process, knowing beforehand how well a classifier will perform on that data can be a very useful insight. The concept of Discernibility aims at exactly that [1], through its metrics. Yet it was only with the creation of the latest index that this insight can be yielded in a very efficient way, making it a viable alternative for feature evaluation among other applications. For this purpose, a number of artificial features, of different class overlap levels were created. These features, along with a typical machine learning benchmark dataset [2] were applied to four different classifiers as well as the proposed metric.

The rest of the paper is structured as follows. In Section 2, a review of the relevant literature is conducted. This is followed by description of the methodology of the introduced metric (Section 3). In Section 4, the experiments related to the aforementioned datasets are described and a discussion of the results is presented. This is followed by the conclusions along with future avenues of research based on this work (Section 5).

2 Literature Review

The concept of Discernibility was formally introduced in previous work of the author [1]. However, this notion has been used even before that, since the idea of class distinguishability has been present in the field of clustering and pattern classification for a while.

In particular, a metric called *Silhouette Width* [3] has been a popular choice for measuring this, in the context of clustering. This measure makes use of inter- and intra-class distances, though it only considers the inter-class distance of the closest class. It has been shown in [1, 4] that it is outperformed by the Spherical and the Harmonic Indexes of Discernibility, which were developed for this particular task. The latter have been tested in a variety of pattern recognition problems with success.

Another metric that performs this task is the *Fisher Discriminant Ratio* [5] which makes use of statistical analysis to evaluate the class overlap. The downside of this measure is that, because of its nature, it only works for one-dimensional data (a single feature). Also, while the other metrics yield a value in a bound interval ([-1, 1] for SW and [0, 1] for SID and HID), FDR may yield any positive value, making its output sometimes difficult to incorporate in larger frameworks, or to compare with other metrics. This shortcoming is addressed in the metric introduced in this paper.

Alternative metrics for this task have been proposed in [6], although they share the same drawback as FDR, as they were designed to evaluate a single feature at a time. Although most of them concentrate on measuring the trend of the features in relation to the class vector (something quite significant for Fault Diagnosis applications), one of them focuses on class distinguishability. This metric, the *Assumption Free Discriminant Ratio*, is basically another version of FDR with the difference that it does not assume any distribution for the data at hand, an approach that is shared by SW, SID and HID as well.

Another metric is that of the Kernel Class Separability method [7], which has application in feature selection [8]. However, this metric this metric is very heavy in terms of parameters, which although they can be fine-tuned, they make this technique quite cumbersome and impractical for real-world problems. In addition, this metric's use in feature selection was tested only using SVMs, a powerful classifier type but a single classifier type nevertheless. Therefore, this metric cannot be considered as a viable alternative for the class separability measurement task, at least not until it is further refined and optimized.

An interesting alternative is presented in [9] where a statistical analysis is performed to evaluate the class separability potential of various features. This is very similar to the FDR technique, though more analytical and therefore computationally expensive. This method has the inherent weakness of the statistical approach, namely its limitation to a single-dimensional dataset, rendering it ideal for feature evaluation but inept for anything more complex.

All of the aforementioned metrics, with the exception of FDR, are to some extent computationally expensive when it comes to larger datasets, a drawback that is addressed by the proposed metric, as it will become evident later on.

3 Methodology

The idea of the metric introduced in this work is to provide a measure of a dataset's Discernibility using inter-class and intra-class distances for each class, for each class pair combination. This is why the metric is called Distance-based Index of Discernibility (DID). Contrary to the Spherical Index of Discernibility, it does not use hyper-spheres, therefore cutting down the computational cost significantly. In addition, it provides only an overall estimate of the Discernibility of the dataset, avoiding the individual Discernibility scores of the comprising data points. This gives it an edge in the computational cost towards both SID and HID, which base their Discernibility estimate on the individual discernibilities of the patters of the dataset.

DID also has the option of using only a sample of the data points, instead of taking the whole dataset. This allows it to tackle datasets with forbiddingly large sizes, without much loss of accuracy in the Discernibility estimation (as it will be seen in Section 4). If the sample size is omitted in the inputs, the whole dataset is used.

Moreover, the DID metric provides the inter-class Discernibility for each class pair, something that, to the best of the author's knowledge, no other similar measure yields as an output. This is particularly useful as it provides insight to the dataset structure, something essential in datasets which due to their dimensional complexity cannot be viewed graphically.

DID's Discernibility estimate is computed by averaging the various inter-class Discernibility scores. The latter are calculated as follows. First the centers of the various classes are found, by averaging all the data points of each class. For example, in a 2-D feature space, if in class A there are 3 points (x_1, y_1), (x_2, y_2) and (x_3, y_3), the center of class A would be $(x_A, y_A) = ((x_1+x_2+x_3)/3, (y_1+y_2+y_3)/3)$.

Then the radius of each class is then calculated, by taking the largest distance from the center of the class to the various class points. In the previous example, if point 2 is the farthest from the center (x_A, y_A), then the radius of class A would be $R_A = \text{sqrt}((x_2-x_A)^2 + (y_2-y_A)^2)$.

Afterwards, a small number is added to all the radii to ensure that there are all non-zero. Then, the sample is divided according to the class proportions and the corresponding amount of data points from each class are selected randomly, for each class. Following that, the distances of these points to each class center are calculated (using the Euclidean distance formula). The ratio of each distance of a pattern of a class *i* over the radius of the class *j* is then computed and adjusted so that it does not surpass the value of 1.

Following that, these ratios are added up for each class pair and then divided by the number of patterns used in the calculations. So if due to the class distributions of the dataset the sample used comprise of *n* patterns in class *i* and *m* patterns in class *j*, the inter-class discernibility of classes *i* and *j* (ICD_{ij}) would be:

$$ICD_{ij} = \frac{\sum_{i=1}^{n} \min(1, dist_{ij} / R_i) + \sum_{j=1}^{m} \min(1, dist_{ij} / R_j)}{n + m} \cdot \quad (\mathbf{1})$$

where ICD_{ij}: inter-class disc. of classes i and j
$dist_{ij}$: distance between patterns i and j
R_i: radius of hypersphere of class i
R_j: radius of hypersphere of class j
n: number of patterns in sample of class i and
m: number of patterns in sample of class j

Note that the above measure is calculated only for different classes ($i \neq j$), for all possible combinations. Moreover, as one would expect, $ICD_{ij} = ICD_{ji}$. So, in a dataset having 4 classes, 6 class combinations will be taken, yielding 6 inter-class distances. As mentioned previously, once the inter-class discernibility scores for each class pair have been calculated, their average yields the overall discernibility score for the whole dataset. So, for a 3-class dataset, DID would be equal to (ICD_{12} + ICD_{13} + ICD_{23}) / 3.

4 Experiments and Results

4.1 Data Description

The data used for this research are a combination of 5 artificial datasets, generated by simple Gaussian distributions, and a benchmark dataset from the UCI machine-learning repository, titled *balance*.

The artificial datasets were designed to manifest five distinct class overlap levels and are single-dimensional (in essence they are features of various quality levels). These datasets comprised of 3000 data points, divided evenly among three classes, which followed a Gaussian distributions with $\sigma = 1$ and various μ's. A typical such dataset is feature3, which exhibits a moderate class overlap, as seen in Fig. 1.

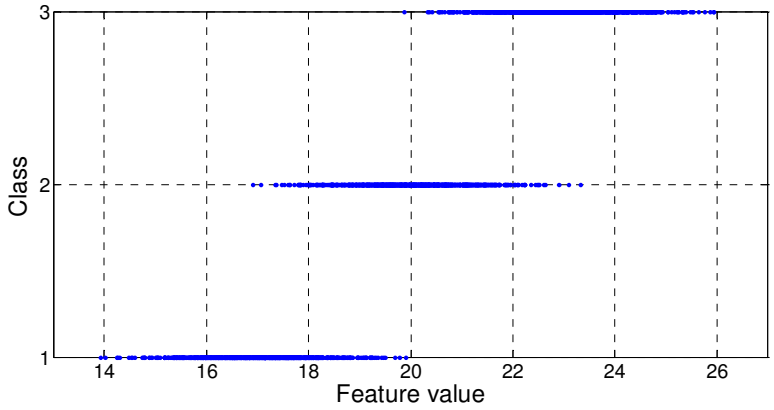

Fig. 1. Mapping of a typical artificial dataset (feature3).

The *balance* dataset was created as part of a cognitive research [11] and comprises of 3 classes as well. It has 625 points of 4 numeric attributes each. The class distribution is unbalanced (288 points in class 1, 49 points in class 2, and 288 points in class 3). The attributes (which are all integers from 1 to 5) describe the following measures: Left-Weight, Left-Distance, Right-Weight and Right-Distance, of a balance scale tip. This dataset has been used mainly for classification research and has no missing values.

4.2 Experiment Setup

A number of experiment rounds over four quite diverse classifiers were carried out. Each round comprised of a 10-fold cross-validation scheme, to ensure an unbiased partitioning of the training and testing sets. The classifiers used were the classic k Nearest Neighbor (using k = 5, which is a popular value for the number of neighbors parameter), the ANFIS neuro-fuzzy system (30 epochs for training), the Linear Discriminant Analysis statistical classifier (LDA) and the C4.5 decision tree. These classifiers were selected because they cover a wide spectrum of classification techniques. Also, in order to ensure more robust results, the number of experiment rounds was set to 30. Most of these classifiers, along with a few others, are thoroughly described in [10].

These experiments were conducted for each dataset and afterwards, the Accuracy Rates of the four classifiers were averaged. The end result was an Accuracy Rate for each dataset, reflecting in a way the classification potential of that data. In addition, each dataset was evaluated using the proposed metric, as well as a few other representative measures: SID, HID, FDR and AFDR. Note that the last two metrics were not applied on the balance dataset as they are limited to one-dimensional data. Also, all of the aforementioned measures were applied on the whole datasets,

although their outputs are not significantly different when applied on the training sets alone (which constituted 90% of the whole datasets, for each classification experiment).

Another set of experiments was conducted in order to perform a sensitivity analysis of the proposed metric and the sample size used. These experiments constituted of 100 rounds and two of the aforementioned datasets were used.

An additional set of experiments was carried out, this time using only the Discernibility metrics, in order to establish a comparison in terms of computational complexity. These experiments comprised of 100 rounds and all of the aforementioned datasets were used.

4.3 Evaluation Criteria

The various outputs of the experiments were evaluated using three evaluation criteria, one for each set of experiments. The relationship between a Discernibility metric with the (average) Accuracy Rate is assessed using the Pearson Correlation (over the six datasets used). For the sensitivity analysis experiments the relative error (in relation to the Discernibility score of the whole dataset) was employed. As for the computational complexity series of experiments, the CPU time measure was used.

4.4 Results

The experiments described previously yielded some interesting results that appear to validate the initial aim of acquiring a reliable insight of a dataset's classification potential, in a way that is computationally inexpensive.

As it can observed from the results of the Accuracy Rates experiments (Table 1), the DID metric appears to follow closely the average Accuracy Rate, for the six datasets used. This close relationship can be more clearly viewed in Figure 2. The correlation coefficient was calculated to be an impressive 99.8%, verifying statistically the above observation.

Table 1. Experimental results for examining the relationship between classification accuracy and DID scores. The accuracy rate is averaged over all four classifiers used and over all thirty rounds.

Dataset	Mean Accuracy Rate	DID score
Feature1	0.9998	1.0000
Feature2	0.9887	0.9921
Feature3	0.8970	0.8909
Feature4	0.5354	0.4265
Feature5	0.3361	0.2468
Balance	0.8116	0.7599

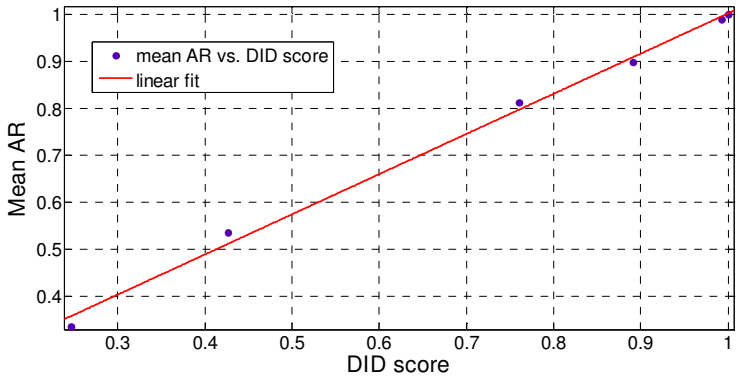

Fig. 2. Relationship between mean Accuracy Rate and DID scores, for all of the datasets tested. It is clear that a vivid (linear) correlation exists.

In order to ensure that the proposed metric is a viable alternative to the existing measures that opt to accomplish the same task, a comparison was made among them. Since a couple of these metrics cannot be applied in multi-dimensional data, two sets of comparisons were made, one using all datasets and one using only the single-dimensional ones (the artificial features created for this research). The results can be viewed in Table 2.

Table 2. Comparison of DID with other Discernibility metrics, based on the (Pearson) correlation with the classification Accuracy Rate, using all 6 datasets and the 5 single-dimensional datasets respectively.

Discernibility Metric	**Correlation w. Mean Accuracy Rate**	
	All datasets	**Only 1-dim datasets**
SID	0.997	0.998
HID	0.976	0.999
DID	0.998	0.998
FDR	–	0.946
AFDR	–	0.971

As the proposed metric has the option of using a sample of the data points in the dataset it is applied on, it is worthwhile investigating how the size of the sample influences the metric's output. This was done in the second set of experiments, which involved two datasets, the balance one and one of the features (feature3). The output of the metric when it is applied using the whole dataset is taken to be the correct Discernibility score, with which all the other outputs are compared (Table 3).

Table 3. Sensitivity analysis of DID scores, over different sample sizes, for two of the datasets used in the classification experiments. The DID scores were calculated over 100 runs. The original DID scores for the two datasets were 0.7599 and 0.8909 respectively.

Dataset	Sample size	Mean	St. Dev.	Rel. Error (%)
Balance	50%	0.7593	0.0060	0.0734
	25%	0.7605	0.0104	0.0790
	12.5%	0.7611	0.0153	0.1535
	5%	0.7558	0.0211	0.5457
Feature3	50%	0.8911	0.0018	0.0197
	25%	0.8911	0.0029	0.0197
	12.5%	0.8911	0.0051	0.0197
	5%	0.8911	0.0076	0.0197

In the third set of experiments the computational cost of the proposed metric, in comparison with the other metrics, is examined. The results of these experiments can be best viewed graphically, as seen in Figure 3 below.

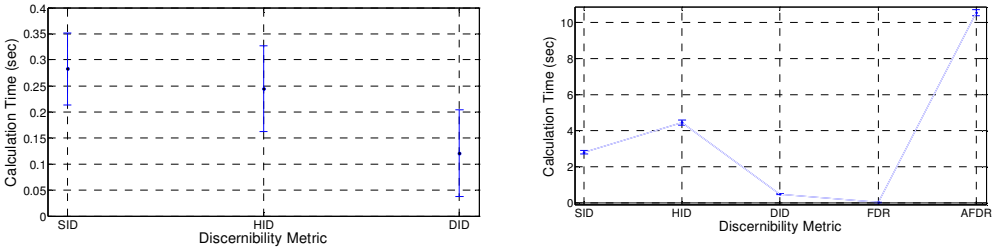

Fig. 3. Computational cost analysis. The calculation time of the Discernibility scores for the various metrics are shown (in sec). Two datasets were used for 100 runs (left = *balance*, right = *feature3*). The error bars depict the 95% confidence intervals of the calculation time.

4.5 Discussion

It is clear from the aforementioned results that the proposed metric maps quite accurately the (average) Accuracy Rate. This in essence makes it a reliable predictor of a classifier's performance, something can translates into a significant advantage in applications where the classification cost is relatively high. Also, it enables the user to have a better understanding of the dataset before the actual classification, something that may help him/her make a better decision regarding the classifier used.

The results of the second series of experiments dictate that the proposed metric is at least as good the other ones, in mapping the Accuracy Rate of the classifiers. Also, in appears to be somewhat better in that aspect, when compared to FDR which has been extensively used in the past for the same purpose.

The results of the sensitivity analysis are quite interesting as they show that the metric's output is not greatly affected by the sample size. As one would expect, the output varies more as the sample becomes smaller, yet the relative error remains quite low (<1%). even at samples of only 1/20th of the original dataset. It is noteworthy that in the case of feature3, where the number of data points is relatively high, the metric's output is quite stable and close to the correct value, even though it varies a bit, as the sample gets smaller.

The computational cost experiments verified the original hypothesis that the proposed metric is an efficient alternative to the other Discernibility measures. When tested on a multi-dimensional dataset against SID and HID, it is clear that it is generally faster. Also, on a single-dimensional data with more data points, the advantage over these two measures is even more evident. It is still not as fast as FDR, but it is significantly faster than AFDR, which is in general a more robust metric than FDR.

Further analysis, using the ROC evalutation criterion for example, could have been performed. However, it is evident from the analysis so far that the proposed metric is adequate regarding the task it undertakes. Besides, a more extensive analysis is beyond the scope of this paper and can be part of a future publication based on further research on the subject.

5 Conclusions and Future Work

From the research conducted it can be concluded that the proposed method is a robust Discernibility metric, yielding a very high correlation with the average Accuracy Rate over the datasets used in the experiments of this research. Apparently it is not as fast as FDR, yet DID provides a better performance that this metric plus it is applicable on multi-dimensional data as well. Also, it has the option of using a sample of the dataset, without deviating much in its output, even for quite small sample sizes.

Future work on this topic will include more extensive testing of the method, in a larger variety of datasets, as well as use of it in other classification-related applications. Also, ways of making it applicable on the data point level will be investigated, so that it can yield Discernibility scores for the individual patterns of the dataset it is applied on. Finally, ways of making use of the inter-class Discernibility assessments of the various class pairs of a dataset will be also explored.

References

1. Z. N. Voulgaris. *Discernibility Concept for Classification Problems.* Ph. D. thesis, the University of London, 2009.
2. UCI repository: http://archive.ics.uci.edu/ml/ datasets.html (last accessed: January 2011).

3. Kaufman, L., Rousseeuw, P. J. Finding groups in data. Wiley Interscience Publications, New York, 83-85, 1990.
4. Z. Voulgaris, B. Mirkin, “Choosing a Discernibility Measure for Reject-Option of Individual and Multiple Classifiers”, International Journal of General Systems, [accepted and pending publication], 2010.
5. Fisher, R. A. The use of Multiple Measurements in Taxonometric Problems and. Eugenics, vol. 7, pp. 179-186. 1936.
6. Z. Voulgaris, C. Sconyers, “A Novel Feature Evaluation Methodology for Fault Diagnosis”. Proceedings of World Congress on Engineering & Computer Science 2010, October 2010, USA, Vol. 1, 31-34.
7. L.Wang, and K. L. Chan, “Learning Kernel Parameters by using Class Separability Measure”. 6th kernel machines workshop, 2002.
8. L. Wang, “Feature Selection with Kernel Class Separability”. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30 (9), pp. 1534-1546, Sept. 2008 (doi:10.1109/TPAMI.2007.70799).
9. Erick Cantú-Paz, “Feature Subset Selection, Class Separability, and Genetic Algorithms”. Lecture Notes in Computer Science, Volume 3102/2004, pp. 959-970, 2004 (DOI: 10.1007/978-3-540-24854-5_96).
10. R.O. Duda, D.G. Stork, and P.E. Hart, Pattern Classification (Second Edition). John Wiley & Sons, 2001.
11. R. S. Siegler, Three Aspects of Cognitive Development. Cognitive Psychology, 8, pp. 481-520, 1976.