Measures for the characterisation of pattern-recognition data sets

Christiaan van der Walt, Etienne Barnard

Human Language Technologies Research Group,
Meraka Institute, CSIR
Department of Electrical, Electronic and Computer Engineering,
University of Pretoria
{cvdwalt,ebarnard}@csir.co.za

Abstract

We study the relationship between the properties of data and classifier performance. Data measures are employed to characterise classification problems and it is shown that these data measures successfully capture important characteristics of the relationship between data and classifiers. The proposed data measures can be used to predict the classification performance of real-world data sets and to gain insight into the structures and properties of real-world data.

1. Introduction

We today have a wide-range of classifiers that are employed in numerous applications, from credit scoring to speech-processing, with great technical and commercial success. No classifier, however, exists that will outperform all other classifiers on all classification tasks and the process of classifier selection is still mainly one of trial and error.

Several empirical studies have shown that the choice of optimal classifier does in fact depend on the data set employed [1, 2], and some guidelines on classifier selection have been proposed [3]. These guidelines do not, however, provide much insight into the specific characteristics of the data that will determine the preference of classifier; several theoretical approaches have also been employed to predict the performance of classifiers in an *a priori* fashion [4–6]. We will show in the next sections that these approaches fall short of a comprehensive solution to the task of classifier selection.

A significant amount of insight into the theoretical properties of classifiers and of data will be required to fully describe the relationship between data characteristics and classifier performance; we will search for such insight by (1) identifying data properties that influence classification performance and (2) measuring these properties from data.

Various experiments have been performed in [7, 8] to identify the properties of data that influence classification performance; in this paper we will propose measures to measure these properties from data. We will also illustrate how these data measures can be used to predict the classification performance of real-world data sets.

In Section 2 we will briefly summarise various approaches that have been proposed in the literature to predict the classification performance of classification tasks; in Section 3 we will propose data measures that capture important characteristics of the relationship between data properties and the performance of classifiers and in Section 4 we will illustrate how these data measures can be employed to construct a meta-classification system. We will explain the results of this meta-classifier in Section 5 and we will conclude on our findings in Section 6.

2. Background

Various strategies have been employed to describe the relationship between classifiers and the problems they try to solve. These approaches are summarised as follows:

- Empirical studies have been performed to compare the performance of classifiers on different real-world data sets [1, 2] and to predict the domain of competence of classifiers [3, 9]. A heuristic meta-learning search method has been proposed by [10] to find the optimal parameters settings of classifiers and to estimate the generalisation performance of these classifiers.
- Data measures to characterise the difficulty of classification problems were studied by [11]; their focus was on the geometrical complexity of the decision boundaries between classes.
- A theoretical framework, known as the no-free-lunch theorems, was developed in [4, 5] to predict and compare the generalization performance of classifiers.
- Statistical learning theories, such as that of Vapnik and Chervonenkis (VC) [6], have been used to place bounds on the generalisation error rates of data sets.

All these approaches are limited in some way or another. The no-free-lunch theorems and the bound on generalisation performance of classifiers using VC dimensions are very limited in terms of real-world applications. Empirical studies have shown the importance of the relationship between data characteristics and classifier performance; they have, however, failed to describe this relationship in detail. A detailed discussion of each of these approaches is given in [8].

3. Data measures

Previous empirical studies have shown that data measures can be employed to give valuable insight into data set properties [11,12]; these studies have, however, failed to explain how these properties influence classification performance.

In this section we will briefly summarise data measures that are specifically designed to measure data properties that influence classifier performance. We will use the data measures proposed in [8]; we will, however, only discuss the most informative data measures, as identified by [8], in this section. A detailed discussion of all the data measures is given in [8].

3.1. Correlation of features

We use the following data measure to measure the average absolute correlation between features in a data set [2]:

$$p = \frac{1}{T} \sum_{i=1}^{C} \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} |p_{jk}|, \quad (1)$$

where $|p_{jk}|$ is the absolute value of the Pearson correlation coefficient between variables j and k and T is the total number of correlation coefficients added together.

3.2. Multivariate normality

We will use the BHEP test for multivariate normality since this test possesses the following desirable properties [13, 14]:

- Affine invariant
- Consistent against non-normal distributions
- Can be applied to data sets of any size and dimensionality.

The calculation of this measure is rather involved; we refer the reader to [13] for a full discussion of this test. We will use this weighted distance measure as a measure of normality and indicate it as *MVN*.

3.3. Linear separability

We use a linear-discriminative classifier described in [15] to perform linear classification. The linear discrimination function is a linear combination of the variables in a sample. This discrimination function is used to construct an optimal hyperplane to discriminate between data of different classes in a d-dimensional feature space. We use the 10-fold cross-validation error rate of this linear classifier as a measure of linear separabiltity. We denote this error rate as $L1$.

3.4. Samples per group

We use an ϵ-neighbourhood pretopology approach proposed by [11, 16], to grow successive adherence subsets from points in each class; each adherence subset is grown to the highest order such that it includes only points of the same class.

The number of samples in the retained adherence subsets gives us an indication of the sizes of groups in the data. The average size of these subsets can be seen as a measure of the average number of samples per group. We use the following measure:

$$T_2 = \frac{1}{N_{retained}} \sum_{i=1}^{N_{retained}} S_i, \quad (2)$$

where $N_{retained}$ is the number of retained adherence subsets and S_i is the number of samples in adherence subset i.

3.5. Variation in feature standard deviation

We will calculate the variation in feature standard deviations (SDs) in each class by calculating the SD of the feature SDs; we use the maximum-likelihood equations given in [8] to calculate these SDs. We denote this SD of the feature SDs as measure $T3$.

3.6. Inter-class scale variation

The scale of data in various parts of the feature space of a data set can be measured by the density of the hyper-spheres retained by the pretopology ϵ-neighbourhoods approach. We define the density of a retained subset as:

$$\rho = \frac{N_{sphere}}{V_{sphere}}, \quad (3)$$

where N_{sphere} is the number of samples in a retained subset and V_{sphere} is the volume of the retained subset. The radius of a sphere is the Euclidean distance from the sphere centre to the furthest sample in the sphere.

We calculate the SD of the sphere densities of a data set to give us an indication of the variation in sphere density in a data set and consequently a measure of variation in scale through the feature space. The SD of sphere densities will give us a measure of both intra-class and inter-class scale variation. We will denote the SD of sphere densities as measure $T4$.

3.7. Input noise

To determine input noise we will determine the amount of overlap between features of different classes; we will follow an approach suggested by [11] with two slight variations - we will rotate the feature axes with an eigenvalue transformation and also consider the number of dimensions in which overlap occurs. The reason for the eigenvalue transformation is to decorrelate the data as much as possible, since correlation can create the false impression that overlap between features exists (when we only consider one feature at a time).

The maximum and minimum values of a feature in each class are used to define boundaries for a feature; if the feature value of a sample lies in the boundaries of another class's feature values then we will assume that this sample contributes to overlap in this specific feature. We will count for each sample in how many dimensions it overlaps and then normalise the total overlap with the product of the number of samples in the data set and the dimensionality of the data set. We will denote this measure of input noise as measure $N1$.

3.8. Feature noise

We use the intrinsic dimensionality measure proposed in [8] to measure the proportion of features that don't contribute to classification. We use the following measure as a measure of feature noise:

$$ID2 = \frac{d - ID}{d} \quad (4)$$

where d is the dimensionality of the data and ID is the intrinsic dimensionality measure.

3.9. Summary of data measures

The data measures discussed in this section are summarised in Table 1 (the efficacy of these data measures are verified in [8]).

Table 1: *Summary of most informative data measures*

Measure	Data property
p	Correlation of features
$N1$	Input noise
$T3$	Variation in feature SD
MVN	Multivariate normality
$L1$	Linear separability
$T4$	Inter-class scale variation
$T2$	Samples per group
$ID2$	Feature noise

Table 2: *Summary of real-world data sets*

Data set	d(Num)	d(Cat)	d	N	C
Iris	4	-	4	150	4
Balance-scale	4	-	4	625	3
Diabetes	4	4	8	768	2
Tic-tac-toe	-	9	9	958	2
Heart	7	6	13	270	2
Australian	6	9	15	690	2
Vehicle	18	-	18	846	4
German	7	13	20	1000	2
Ionosphere	34	-	34	351	2
Sonar	60	-	60	208	2

4. Meta-classification

In this section we will construct a meta-classification system to predict the classification performance of real-world data sets. We will make use of the data measures proposed in [8] to characterise data sets; these measures include the measures discussed in the previous section. We will also utilise artificial data sets to construct a meta-classifier.

4.1. Classifiers

We will use model-based and discriminative classifiers to construct our meta-classifier; these classifiers are the Naïve Bayes (NB), Gaussian (Gauss), Gaussian Mixture Model (GMM), Decision Tree (DT), k-Nearest Neighbour (kNN), Multilayer Perceptron (MLP) and Support Vector Machine (SVM) classifiers. More details regarding these classifiers are given in [8]. The meta-classifier will be able to predict the performance of these classifiers for real-world data sets.

4.2. Artificial data

We will make use of artificial data sets to construct a meta-classification training set; these artificial data sets are generated with very specific data properties that influence classification performance. Artificial data sets are generated by sampling from a prescribed density function. These density functions are the uniform, Gaussian, Gaussian mixture model and Cauchy probability density functions.

More information on the generation of these artificial data sets and their properties is given in [8].

4.3. Real-world data sets

We will predict the classification performances of ten real-world data sets obtained from the UCI Machine Learning repository [17]. These data sets are summarised in Table 2. We abbreviate dimensionality as d, number of samples as N and number of classes as C. The number of numerical attributes is abbreviated as $d(Num)$ and the number of categorical attributes as $d(Cat)$. The Diabetes, Heart, Australian, Vehicle and German data sets were previously studied in the Statlog project [2].

4.4. Data characteristics

We will make use of the following data characteristics and their corresponding data measures to characterise a data set: correlation of features, multivariate normality of class conditional probability density functions, variation in feature SDs, sparsity of data, input noise, output noise, intra-class scale variation, inter-class scale variation, variation in decision boundary complexity, intrinsic dimensionality, groups per class, samples per group and the interleaving of groups of different classes.

The most informative data measures have been discussed in Section 3 and a detailed description of all these data measures is given in [8].

4.5. Meta-classifier

The flow diagram in Figure 1 illustrates the process used to predict and evaluate the classification performance of real-world data sets.

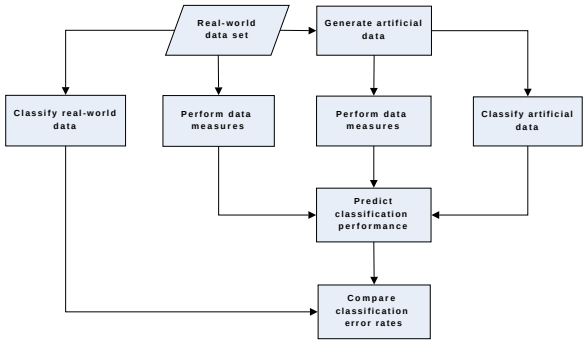

Figure 1: *Flow diagram of meta-classification system*

The data measures proposed in [8] are employed on a real-world data set and artificial data sets are generated with the exact same dimensionality, number of samples and number of classes; these artificial data sets contain various data properties that were identified in [8]. Data measures are employed on these artificial data sets and the 10-fold cross-validation classification error rates of the artificial data sets are determined.

A weighted Euclidean distance is used to compare the data measures of the real-world data set to the data measures of the artificial data sets. The artificial data set closest to the real-world data set (in terms of Euclidean distance) is considered as the data set with the most similar data properties. The classification error rates of this artificial data set are used as the predicted error rates of the real-world data set.

The classification error rates of the real-world data set are estimated by performing 10-fold cross-validation; these error rates are used to evaluate the efficacy of the meta-classifier by comparing them to the predicted classification error rates.

The accuracy of the meta-classifier predictions for all the real-world data sets are given and discussed in the next section.

Table 3: *Classification error rates of real-world data sets*

Data set	NB	Gauss	GMMd	GMMf	kNN	DT	SVM	MLP
Iris	0.0467	0.0200	0.0400	0.0333	0.0333	0.0600	0.0267	0.0400
Balance-s.	0.0960	0.0983	0.2720	0.0832	0.0976	0.2176	0.0000	0.0512
Diabetes	0.2422	0.2579	0.2566	0.2695	0.2500	0.2630	0.2305	0.2227
Tic-tac-toe	0.2265	0.3011	25.00	0.2140	0.0313	0.0438	0.0939	0.0167
Heart	0.1667	0.1704	0.1519	0.1814	0.1926	0.2037	0.1519	0.1667
Australian	0.2290	0.2103	0.1942	0.2029	0.1478	0.1507	0.1464	0.1217
Vehicle	0.5627	0.1451	0.5638	0.1525	0.2943	0.2731	0.1478	0.1690
German	0.2510	0.2890	0.3200	0.3220	0.2690	0.2600	0.2120	0.2490
Ionosphere	0.1738	0.0765	0.3589	0.3049	0.1311	0.1168	0.0884	0.0855
Sonar	0.3173	0.3500	0.1680	0.3269	0.1490	0.2933	0.2260	0.1490

Table 4: *Predicted error rates of real-world data sets*

Data set	NB	Gauss	GMMd	GMMf	kNN	DT	SVM	MLP
Iris	0.1867	0.0400	0.1533	0.0400	0.1133	0.1667	0.0667	0.0667
Balance-s.	0.6333	0.6400	0.6083	0.6250	0.6117	0.6700	0.6250	0.6117
Diabetes	0.4838	0.5275	0.4925	0.5175	0.4975	0.5050	0.4788	0.4888
Tic-tac-toe	0.3052	0.3156	0.1885	0.1865	0.2063	0.1323	0.2531	0.1479
Heart	0.4500	0.4150	0.4100	0.3850	0.4550	0.5100	0.4350	0.4150
Australian	0.3786	0.4757	0.3271	0.2443	0.1414	0.3757	0.2671	0.3086
Vehicle	0.0248	0.0000	0.0248	0.0000	0.0142	0.1168	0.02005	0.0224
German	0.3600	0.4330	0.3560	0.3740	0.3690	0.4110	0.3390	0.3420
Ionosphere	0.1893	0.4567	0.1864	0.4522	0.4096	0.1808	0.1525	0.2175
Sonar	0.0150	0.3550	0.0000	0.0050	0.0000	0.1350	0.0000	0.0050

5. Results

The classification results of the ten real-world data sets are given in Table 3. We will compare these error rates to the predicted error rates of the meta-classifier in order to evaluate the performance of the meta-classifier; the predicted error rates are given in Table 4.

The Pearson correlation coefficients between the 10-fold cross-validation classification error rates and predicted classification error rates are calculated for each data set. These correlation coefficients give us an indication of how accurately the measurements can explain the behaviour of the classifiers. The classifier correlation coefficients are given in Figure 2.

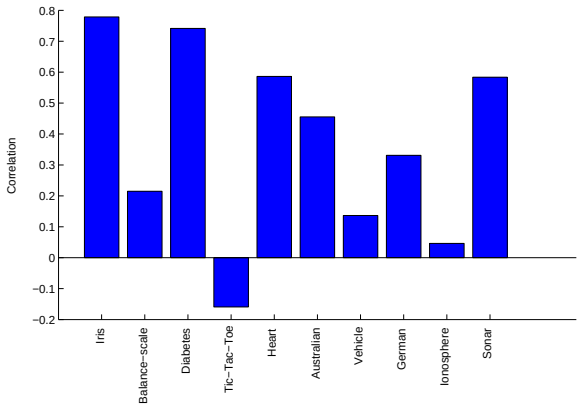

Figure 2: *Correlation coefficients of real-world data sets*

We see in Figure 2 that the Tic-tac-toe data set is the only one with a negative correlation coefficient; this is to be expected, since this data set is the only one that contains only categorical features. Closer evaluation of the classification error rates in Table 3 reveal that the NB, Gaussian, GMMd and GMMf classifiers have very poor classification performance for this data set; this is due to the fact that these classifiers are not suited for categorical data. All the data sets in the meta-training set contain continuous variables, which explains why the predictions of these error rates are not accurate.

The Ionosphere data set has the lowest correlation coefficient of the non-categorical data sets. If we investigate the classification error rates and the predicted classification error rates more closely we find that the predicted error rate of the Gaussian classifier differs significantly from the 10-fold cross validation error rate. If we calculate the correlation coefficient excluding the Gaussian classifier we obtain a correlation coefficient of 0.2861.

The two data sets with the highest correlation coefficients are the Iris and Diabetes data sets. The artificial data set nearest to the Iris data set has Gaussian distributed classes with feature SDs close to unity; the nearest data set to the Diabetes has GMM distributed classes with 100 groups per class with feature SDs between 0 and 5.

What is interesting is that the Diabetes data set contains four numerical and four categorical features; if we evaluate the classification error rates we observe that these four categorical features do not influence the model-based classifiers too negatively compared to the discriminative classifiers. This explains why the correlation coefficient is still very good even though the

data set contains categorical attributes.

The remaining data sets have correlation coefficients between 0.1364 and 0.5862; these data sets and their data measures are discussed in more detail in [8]. These results show that important data characteristics are captured by the employed data measures.

We calculate the correlation coefficients for each classifier across the ten real-world data sets to give us an indication of how well the data measures describe the properties of each classifier; the Pearson correlation coefficients between the classification error rates of each classifier are given in Figure 3.

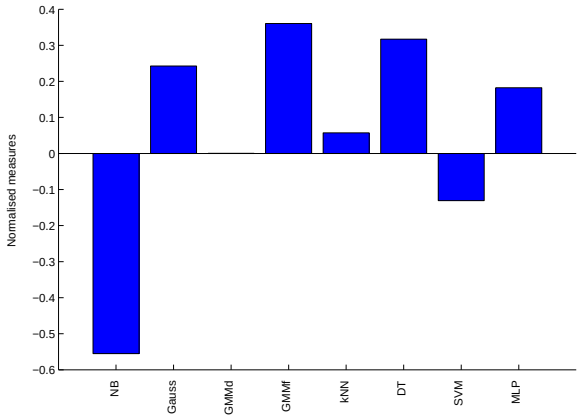

Figure 3: *Correlation coefficients of classifiers*

We see that the NB and SVM classifiers have negative correlation coefficients across all the real-world data sets, and that the observed correlation values are generally lower than the values across the different classifiers for a fixed data set. This suggests that our features are more successful in predicting the relative performance of different classifiers across the same data set than error rates across data sets. This is not surprising in light of the tremendous variability of data sets. Fortunately, the prediction of relative classifier performance is also the more interesting task from a practical perspective.

6. Conclusion

Understanding the relationship between classifiers and the problems they try to solve is crucial in selecting the optimal classifier for a classification task. In this paper we have identified informative data measures that capture some of the characteristics of this relationship.

We have illustrated how these data measures can be employed to characterise a data set and how these data measures can be used to predict the classification performance of real-world data.

The data measures used to characterise a data set allows us to quantify important data properties such as correlation of features, multivariate normality of class conditional probability density functions, variation in feature SDs, sparsity of data, input noise, output noise, intra-class scale variation, inter-class scale variation, variation in decision boundary complexity, intrinsic dimensionality, groups per class, samples per group and the interleaving of groups of different classes.

Positive correlation coefficients were obtained between the true and predicted classification error rates of all the non-categorical real-world data sets. These results show that the meta-classifier captured important characteristics of the relationship between data and classifier performance.

The performance of the meta-classifier across all real-world data sets for each classifier, however, suggests that further insight into the properties of data is required to fully describe the relationship between data characteristics and classifier performance.

7. References

[1] D.M.J. Tax and R.P.W. Duin, "Characterizing one-class datasets," *in Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pp. 21–26, 2005.

[2] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, *Machine learning, neural and statistical classification*, Ellis Horwood Limited, Hemel Hempstead, 1994.

[3] P.B. Brazdil, J. Gama, and B. Henery, "Characterizing the applicability of classification algorithms using meta-level learning," *in Proceedings of the European Conference on Machine Learning*, vol. 784, pp. 83–102, 1994.

[4] D.H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, no. 7, pp. 1341–1390, 1996.

[5] D.H. Wolpert, "The bayesian and computational learning theories," *NASA Ames Research Center, CA, MS 269-1*, Oct. 2000.

[6] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[7] C.M. van der Walt and E. Barnard, "Data characterisitcs that determine classifier performance," *in Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pp. 160–165, 2006.

[8] C.M. van der Walt, "Data measures that characterise classification problems," *Master's dissertation, Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa*, 2007.

[9] B. Pfahringer, H. Bensusan, and C. Giraud-Carrier, "Meta-learning by landmarking various learning algorithms," *in Proceedings of the Seventeenth International Conference on Machine Learning*, vol. 951, no. 2000, pp. 743–750, 2000.

[10] A. van den Bosch, "Wrapped progressive sampling search for optimizing learning algorithm parameters," *in Proceedings of the Sixteenth Belgian-Dutch Conference on Artificial Intelligence*, pp. 219–226, 2004. [Online]. Available: http://ilk.uvt.nl/ antalb/paramsearch. [Accessed: August 24, 2007].

[11] T.K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 289–300, 2002.

[12] R. Engels and C. Theusinger, "Using a data metric for preprocessing advice for data mining applications," *in Proceedings of the European Conference on Artificial Intelligence*, pp. 430–434, 1998.

[13] N. Henze and T. Wagner, "A new approach to the bhep tests for multivariate normality," *Journal of Multivariate Analysis*, vol. 62, no. 1, pp. 1–23, 1997.

[14] G.J. Szkely and M.L. Rizzo, "A new test for multivariate normality," *Journal of Multivariate Analysis*, vol. 93, no. 1, pp. 58–80, 2005.

[15] A.R. Webb, *Statistical Pattern Recognition*, John Wiley, NJ, 2nd edition, 2000.

[16] F. Lebourgeois and H. Emptoz, "Pretopological approach for supervised learning," *Proceedings of the Thirteenth International Conference on Pattern Recognition*, pp. 256–260, 1996.

[17] C.L. Blake and C.J. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://mlearn.ics.uci.edu/MLRepository. [Accessed: August 24, 2007].