

# CANDLE: Classification And Noise Detection With Local Embedding Approximations

Erik Thordsen<sup>1</sup>, Erich Schubert<sup>1</sup>

<sup>1</sup>TU Dortmund University, Dortmund, Germany

## Abstract

The machine learning tasks of supervised classification and unsupervised noise detection are commonly performed separately. In this paper, we propose a combination of both tasks that is based on a score of how close a sample is to the manifold spun by the training data. This implicitly learns the manifold structure of each class. The resulting classifier achieves good accuracy on clear decisions but struggles with overlapping regions that can be excluded from the classification. The performance of this approach is discussed on artificial and natural data sets, and the relationship to intrinsic dimensionality is discussed.

## Keywords

Classification, Outlier Detection, Noise Detection, Manifolds

## 1. Introduction

The goal of classification is to distinguish between two or more groups of objects created by different mechanisms based on the observed features. When the features are numeric, it is commonly assumed that the observations span a vector space in which the underlying mechanism of each class creates points within a confined subspace that can be interpreted as a parametrically bounded manifold. Consequentially, classification approaches like Neural Networks, Gaussian Mixture Modelling, Naïve Bayes variants, and Support Vector Machines exploit this concept to some extent by mimicking these manifolds or their pairwise boundaries. Similarly, approaches from the field of manifold learning attempt to approximate the generative function of the manifold to reduce the dimensionality of the data, ideally to the intrinsic dimensionality (ID) of the manifold, i.e., the number of parameters required to describe the manifold, or even lower. The reduced data can then be used in machine learning algorithms with reduced complexity and a better analogy of distance metrics and semantics. Estimating the required number of parameters of the manifold, the ID, is a research field in itself with primarily two main branches: Estimators either analyze the geometry of the dataset or the expansion rate of distances between points. The geometry-based approaches approximate properties of the implicitly assumed function mapping the parameter space to the observed feature space whereas expansion-based approaches assume some (typically uniform) distribution and analyze the increase in distances to the neighbors of a point.

---


LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany

✉ erik.thordsen@tu-dortmund.de (E. Thordsen); erich.schubert@tu-dortmund.de (E. Schubert)

🆔 0000-0003-1639-3534 (E. Thordsen); 0000-0001-9143-4880 (E. Schubert)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In this paper, we propose a combination of manifold learning and classification that skips the explicit modeling of the manifold and rather computes distances to the observed manifold structure. We use the  $k$ -nearest-neighbor distance with a twist: Each training sample uses its own Mahalanobis distance. Thereby, we additionally obtain a measure of certainty of a query point belonging to some class by interpreting the distance as a multivariate deviation. By rescaling these deviations we can account for different sampling densities between different classes, making the certainty measures comparable. In the same manner, this approach could be extended to an outlier-score-based classification scheme.

We first give an overview of related work in Section 2. In Section 3 we argue why the classifier properly describes the manifold structure of the training data, why we expect the classifier to both distinguish between clear and close classification decisions, and how to obtain noise labels. An experimental evaluation is provided in Section 4 followed by a conclusion in Section 5.

## 2. Related Work

At the heart of our approach are three concepts: Classification, outlier or noise detection, and manifold learning. Each of these concepts by itself has numerous implementations on which we will not elaborate here. To our knowledge, this is the first approach to actively include all three concepts in one method. In the remainder of this section, we will, hence, discuss methods with pairwise combinations of these concepts, which are related and of interest for this paper.

The combination of classification and outlier detection has been examined by, e.g., Guan and Tibshirani in their BCOPS algorithm [1] and related work like the Density Level Sets by Chen, Genovese and Wassermann [2]. Similar to our goals, these approaches distinguish between clear classifications, instances that could be contained in multiple classes, and instances that likely belong to neither class. However, these approaches solely focus on the case of binary classification. The BCOPS algorithm further builds on another classifier to enhance it with noise detection capabilities instead of being a classifier in itself. In contrast, our approach starts with a manifold-based noise detection and expands it to a classification model.

The concepts of classification and manifold learning are a classical combination in machine learning. Manifold learning can be used for feature reduction prior to classification with any classifier by applying, e.g., Locally Linear Embedding [3], or t-SNE [4]. Secondly, classifiers can actively use the manifold assumption for classes to fit a function to their structure. To some extent, one can argue, that neural networks with one-hot encoding employ this combination. The decision boundary drawn by Kernel-SVMs could – dependent on the kernel – also be accounted to the manifold structure of classes. There, however, exist extensions of SVMs to actively include the manifold assumption [5]. The most similar work to our classifier that we could find is by Li and Dunson [6] who use a different approach to approximate the local manifold structure. Yet, they do not use the quality of fitness to the manifolds to also decide upon close decisions and noise.

The combination of outlier detection and manifold learning has been studied in works on outlier detection like, e.g., the Correlation Outlier Probability (COP) by Kriegel et al. [7] or the Angle-Based Outlier Detection by Kriegel et al. [8]. The concept to use the geometric shape of the data to tell signal from noise is a nearby idea when thinking about outlier detection. Most

closely related to our approach is COP as it also explains deviation from the manifold in terms of local distribution observations. To our knowledge, no manifold-based outlier detection method has actively been used as a classification model. Our approach could, however, also be adapted to these methods. Manifolds and local correlations have also been considered in cluster analysis, e.g., by the methods 4C [9], COPAC [10], LMCLUS [11], ERiC [12], and CASH [13]. Some only consider clusters that have a linear shape, others measure the deviation in an arbitrarily oriented subspace by selecting principal components.

The complexity and shape of manifolds in data sets have also been studied in the field of intrinsic dimensionality estimation which focuses on estimating the number of parameters of the generative mechanism producing the manifold. While global approaches like PCA can be used, more recent methods rely on local estimates (i.e., per-point) to account for non-linear embeddings and varying complexity throughout the data set. Methods in the ID estimation field generally fall into two major categories: Distance-based approaches like the MLE [14], GED [15], ALID [16], or TLE [17] estimators base their estimates on the speed at which distances to larger neighborhoods increase. Geometry-based approaches like the local PCA, FCI [18], or ABID [19] estimator base their estimates on the geometric shape of neighborhoods. Distance-based ID estimates have successfully been used, e.g., in applications for spatial search [20], clustering [21], and outlier detection [22]. Although the noise definition in this paper is not directly derived from ID estimates, we use the manifold approximation techniques employed in the geometrical ID estimation motivated MESS framework [23] for supersampling data sets.

### 3. From Geometric Noise To Classification

Our approach aims to answer the following question: Does a query point lie on the manifold describing a class and if not, how far away from the manifold is the point? We use the approximation of the Jacobian of the implicitly posited embedding function introduced in MESS [23]. The idea of this approach is, that we can approximate the embedding function describing the class manifold by using local covariance matrices. We can then interpret the Mahalanobis distance as an approximation of the Euclidean distance in parameter space, i.e., the preimage of the embedding function, with an exaggerated additive term for components orthogonal to the embedding function. Ideally, the local covariance matrices should be singular when the data truly lies on a lower-dimensional manifold. Practically (and by adding a very small constant to the diagonal of the covariance matrices) this is rarely the case, giving a set of close-to-zero eigenvalues. Points in the direction of the corresponding eigenvectors suffer a large penalty in scale reciprocal to these eigenvalues. The smaller the approximation errors due to curvature of the manifold and measuring noise, the larger the penalty for points lying outside the manifold in the Mahalanobis distance. This contrasts earlier approaches in cluster analysis (e.g., 4C [9] and COPAC [10]) as well as in outlier detection (e.g., COP [7]) as we do not select directions based on a threshold on the eigenvalues, nor actively put a penalty on small eigenvalue directions. We instead use the regular Mahalanobis distance. Following the MESS framework and the ID estimation literature, we use point distributions around our points of interest rather than the mean of their neighborhoods. For that, we use a non-standard modification of the covariance matrix, discussed below.

Given a point of interest  $x$  and a set of vectors  $x_1, \dots, x_n$  (which does not need to contain  $x$ ), the non-centered covariance matrix regarding the point of interest  $x$  is defined as

$$\text{Cov}(X, x) = \frac{1}{n}(X - x)^T(X - x) \quad (1)$$

where  $X \in \mathbb{R}^{n \times d}$  is a matrix with the vectors  $x_i$  as row vectors and  $X - x$  describes the row-wise subtraction. The resulting matrix is similar to the classical sample covariance matrix (if  $x$  were the arithmetic mean) but more closely describes the geometry around the point of interest  $x$ .

By generating a non-centered covariance matrix for each point in the training data, using the  $n$ -nearest neighbors as per Euclidean distance of  $x$  from its class  $C$  (written as  $\mathcal{N}_n(C, x)$ ), we can define an individual Mahalanobis distance for each point  $x$  as (to avoid singular matrices, we add a small  $\varepsilon$  to the diagonal prior to matrix inversion)

$$d_x(q) = \sqrt{(q - x)^T (\text{Cov}(\mathcal{N}_n(C, x), x) + \varepsilon I)^{-1} (q - x)}. \quad (2)$$

As the distances are large for points outside the manifold and small on the manifold of each class, we can use the  $k$ -nearest-neighbor-distance ( $k$ -distance) to decide if a query point is similar to some class. Combined with the individual Mahalanobis distances  $d_x$  we obtain a score giving low values to query points close to the locally linear continuation of the class manifold of at least  $k$  points from the class. We proceed to compute the mean and standard deviation of the  $k$ -distances of all training samples of each class to compute a score of how likely a query point lies on the class manifold. We define the plausibility score  $L(q, C, k)$  that describes how likely a query point  $q$  is inside a class of training samples  $C$  considering the  $k$ -distances as

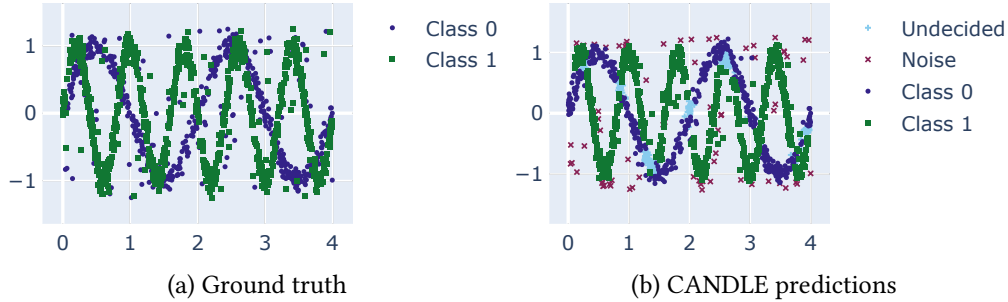
$$L(q, C, k) = \max \left( 0, \min \left( 1, \frac{d_k(q, C) - \mu_{x \in C}(d_k(x, C))}{c \cdot \sigma_{x \in C}(d_k(x, C))} \right) \right) \quad (3)$$

where  $d_k(q, C)$  is the  $k$ -smallest  $d_x(q)$  of all  $x \in C$  and  $\mu_{x \in C}(d_k(x, C))$  and  $\sigma_{x \in C}(d_k(x, C))$  are the mean and standard deviation of  $k$ -distances of all  $x \in C$ , respectively. The cut-off value  $c$  is a user-defined parameter to control how “conservative” the model should be, as the plausibility score drops to 0 beyond  $c$  standard deviations above the mean  $k$ -distance. This way, all points orthogonal to the shape of each class as well as points that lie alongside some class yet outside the observed subspace described by the training data can be discriminated.

Given training samples  $C_1, \dots, C_m$  from a set of  $m$  classes, we can then classify a query point  $q$  based on the function

$$f(q, k) = \begin{cases} \text{noise} & \text{if } \forall i : L(q, C_i, k) = 0 \\ \text{undecided} & \text{if } \nexists i \forall j \neq i : L(q, C_i, k) \geq b + L(q, C_j, k) \\ & \text{and } \exists i, j \neq i : L(q, C_i, k), L(q, C_j, k) > 0 \\ \text{argmax}_{C_i} L(q, C_i, k) & \text{otherwise} \end{cases} \quad (4)$$

where  $b$  is a user-defined parameter that defines how confident the classifier needs to be to distinguish between two classes. This allows us to not label points that lie at the intersection of two classes. If  $b = 0$  then no point will be labeled undecided.



**Figure 1:** Classification of a training data set consisting of two sine waves with 600 and 1000 points, respectively, using CANDLE parameters  $n = 20, k = 15, c = 1, b = 0.1$ , and  $\varepsilon = 10^{-8}$ . A demo implementation to test the CANDLE classifier and create a similar plot can be found at <https://github.com/eth42/candle-demo>.

The resulting method is named **CANDLE (Classification And Noise Detection With Local Embedding Approximations)**, and the user has to choose these parameters:

- $n$  The number of neighbors used for covariance matrix computation. Similarly to neighborhood sizes in ID estimation [14, 19], we propose values at least proportional to the square of the ID, e.g., three times the squared ID. Larger values oftentimes give better results as long as the neighborhoods can still be considered locally linear.
- $k$  The number of neighbors used in the plausibility score computation. This value behaves similar to the  $k$  in the classical  $k$ -nearest neighbor classifier, except decisions between, e.g., two classes use twice the number of reference points in CANDLE compared to the kNN classifier.
- $c$  Multiple of standard deviations necessary to label a query point as noise.
- $b$  The minimum difference in plausibility score for how distinct classification labels need to be to not be undecided.
- $\varepsilon$  To ensure non-singular covariance matrices. This value needs to be orders of magnitude smaller than distances between training samples but can be chosen mostly arbitrarily.

Figure 1 shows an example of how the classifier labels two noisy sine waves. Most noise samples that do not lie on the curve are labeled as noise and the classifier marks the intersections of the manifolds as undecided. In practice, this approach, hence, allows to classify most query points automatically while actively assigning noise- and undecided-labels to query points that would likely be mislabeled. This makes the approach resilient to adversarial attacks and reduces the number of labeling errors on non-noise and non-undecided query points. In case no noise- or undecided-labels are desired, the function  $f(q, k)$  can be replaced by an argmax on the  $L(q, C, k)$  scores, possibly even discarding the min and max in its definition. The scores  $L(q, C, k)$  can also be used immediately as a type of soft classification.

Naively implemented, the computational cost during classification is quite high, as the  $k$ -distance using a different distance function for each training sample can not trivially be sped up with common index structures. As the covariance matrices of nearby training samples should differ only to a small extent, creating a search tree may be possible, but we do not have an implementation, yet. A simple approach to reduce the computation time with very large

**Table 1:** Statistics of the data sets used. The mean ABID estimates were computed for the same neighborhood sizes as the covariance matrices in CANDLE.

Data set	Instances	Features	Classes	Smallest class	Biggest class	mean ABID
Strips	3307	2	5	485	799	1.86
NBA	8536	38	5	970	3030	4.73
MNIST-PCA	70000	50	10	6824	7877	6.19

training sets is to use only a subset of the training data during classification. The covariance matrices are still computed on the  $n$ -nearest neighbors in the full training set, but only a subset is used for classification, which makes the brute force search for the  $k$ -distances faster.

The CANDLE approach shifts the goal of classification from finding the best fitting class to computing a fitting score for each class and assigning labels based on the confidence of fit – much like probabilistic algorithms such as Gaussian Mixture Modelling. Instead of finding a good representation of the decision boundary between classes like decision trees or support vector machines, the CANDLE approach attempts to find a good representation of the decision region of each class, if anything similar to neural networks with one-hot encoding. This can naturally be applied to any fitting score like outlier scores. The Mahalanobis distance-based approach assigns values approximately proportional to deviations of the training data and, thereby, gives comparable values across different classes.

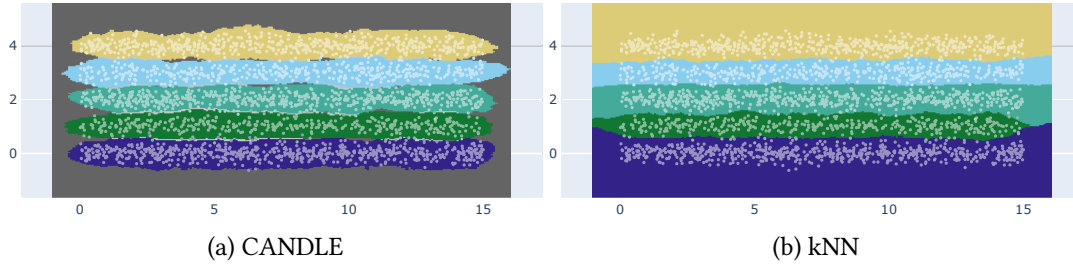
## 4. Experiments

In our experiments, we use both artificial and natural data sets. Table 1 gives some statistics on the data set properties. Additionally to common metrics such as size, we give the mean ABID [19] estimates of intrinsic dimensionality for each of the data sets to give an overview of the intrinsic complexity of the data sets. The ID estimates are averaged over the entire data set, not computed per class. While they could vary between classes (and for overlapping regions), they nevertheless indicate some kind of data set complexity.

For the reference classifiers, we used the scikit-learn implementations. Unless stated otherwise, we evaluated all recall and accuracy values on a 20% holdout test set and optimized model parameters using grid search and 5-fold cross-validation on the remaining 80% using accuracy as the target score. We always use  $\varepsilon = 10^{-8}$  to avoid singularities, as this scale was appropriate for all data sets. All recall and accuracy values are presented in percentages.

### 4.1. Strips Data

The strips data set consists of 5 noisy lines with varying sample counts parallel to each other. Figure 2 displays the decision regions for both a CANDLE model and a  $k$ -nearest neighbors model. The  $k$  for the  $k$ -nearest neighbors model was chosen twice the  $k$  of the CANDLE model to have the same number of samples involved in labeling decisions between two classes. As the CANDLE approach is covariance-based and uses the distance distribution of the training samples, it treats all classes with equal weight, whereas the  $k$ -nearest neighbor model prefers the more frequent classes. This can be seen by the dark green (second from below) strip, which is the



**Figure 2:** Decision regions of a CANDLE model with  $n = 64$ ,  $k = 16$ ,  $c = 3$ , and  $b = 0.15$  and a  $k$ -neighbors model with  $k = 32$ . The regions labeled with class labels are colored according to the overlaid training samples. Regions labeled noise are dark grey and the light regions in between classes correspond to undecided labels.

same width as the strips above and below in the CANDLE plot even though the corresponding class has about 25% fewer points than the other two classes. In the  $k$ -nearest neighbor plot, the dark green strip is partially taken over by the other two strips. Especially the undecided regions are preferably assigned to the more frequent class by the  $k$ -nearest neighbor approach.

Figure 2 also shows how the decision regions expand with the covariance. While the decision regions end close to the training samples vertically, they are slightly elongated horizontally. This corresponds to the intuition that points in the direction of the strip are more likely to belong to the class than points above or below a strip. Larger values of  $c$  naturally increase the tails left and right while larger  $b$  values increase the regions between classes deemed undecided. Increasing  $n$  smoothes the decision boundary as the covariance matrices are less dependent on local sampling densities. For curved manifolds, increasing  $n$  too much might, however, introduce unwanted orthogonal components and compromise the quality. Increasing  $k$  has a similar result by evening out discrepancies due to larger neighborhoods.

## 4.2. NBA

The NBA data set provided at kaggle<sup>1</sup> contains information about both players and games from the NBA dating back to 1950. From this, we created a data set with one entry per player and season for the seasons from 2000 to 2015. We ignored all player-season combinations where the player did not play in at least 5 games or played less than a total of 15 minutes in that season to reduce the amount of noise in the data set. For each player and season, we used the player statistics like height, weight, age in that season, goals, efficiency scores, free throws, and so on. All values that naturally correlate with playing time, like goals, were divided by the minutes played in that season. All features not directly indicative of player performance like the number of games and minutes played as well as all non-numeric features were dropped. The resulting features were then standardized. The possible player positions “C” (center), “F” (forward), “G” (guard), “C-F”, “F-C”, “F-G”, and “G-F” were used as class labels, although we combine “C-F” and “F-C” respectively “F-G” and “G-F” to one label. We fixed the CANDLE parameters to  $c = 2$  and  $b = 0.05$  to control how conservative the classifier should act. The selected model

<sup>1</sup><https://www.kaggle.com/drgilermo/nba-players-stats>

**Table 2:** Accuracy (A) on either all, non-noise/undecided-labeled, noise-labeled, and undecided-labeled samples of the NBA test set. Additionally, “relaxed” accuracy (RA) values are given that also accept overlapping labels like “C-F” for “C” and vice versa as correct. The best mean accuracy and “relaxed” accuracy values for each row are highlighted.

Class	CANDLE				kNN		Random Forest	
	Noise	Undec.	A	RA	A	RA	A	RA
All	1.82	23.14	(53.71)	(72.99)	68.54	88.93	<b>80.32</b>	<b>94.79</b>
Non-Noise/Undec.	-	-	71.58	<b>97.27</b>	71.66	91.33	<b>82.44</b>	96.17
Noise	100.00	-	-	-	51.61	93.55	<b>93.55</b>	<b>100.00</b>
Undec.	-	100.00	-	-	59.75	80.76	<b>72.41</b>	<b>89.87</b>

parameters then were CANDLE  $n=16$ ,  $k=16$ ,  $c=2$ , and  $b=0.05$  for CANDLE,  $k=64$  for the  $k$ -nearest neighbor classifier and 100 trees for the random forest classifier. The results of the experiment are displayed in Table 2 and Table 3. Aside from giving accuracy and recall values, we added “relaxed” accuracy and recall values, that accepted all “similar” class labels as correctly classified. For the unary classes (C, F, G), the relaxed values also accept any combinations containing these values as correct classification. For the binary classes (C-F, F-G), the relaxed values also accept the two parts as correct classification. By allowing these semantically very similar classes to be accepted as correct labelings, we intended to reduce the error due to these classes overlapping. As can be seen from the results, the classes are highly overlapping with up to 34.96% of each class being labeled undecided by the CANDLE model even though the parameter  $b$  is relatively small. This can easily be explained by, e.g., the player roles “C” and “C-F” not being cleanly separable, as the latter is commonly defined as players playing both center and forward roles frequently. A player that is solely a center player and a player that sometimes also acts as a forward will have a similar play style and scoring statistics. When also accepting overlapping classifications as correct, we can see the “relaxed” recall values being much better than the pure recall values. In comparison to other classifiers like the highly related  $k$ -nearest neighbor classifier, the CANDLE approach gives comparable if not better results when reducing the amount of overlap. In practice, this means, that by increasing the parameter  $b$  which controls how much undecided labels are assigned, the accuracy of non-noise and non-undecided labels can be increased in the case of highly overlapping classes.

The accuracy values of CANDLE on the entire data set are decreased due to the noise and undecided labels, as these are considered mislabeled. We, hence, put these values into brackets and provide an additional row considering only instances that are neither labeled noise nor undecided to give an accuracy score for the classified instances only in Table 2. As it would otherwise be unfair for CANDLE to “cherry-pick” which instances to classify, we added accuracy values for the same filtered instances for the reference classifiers as well. As adding a separate row for each class in Table 3 would be cumbersome to read, we combined the unfiltered and filtered recall values for each class into one row, signaling the filtering with a \*-symbol. The unfiltered values for CANDLE are omitted due to redundancy (product of recall values and percentage of classified instances).

Yet, the improvement due to ignoring the noise and undecided labeled instances does not only apply to CANDLE. Removing the test samples that were labeled either noise or undecided,



**Table 3:** Recall (R) of individual classes on the NBA test set. Columns marked with \* are computed on all instances that are labeled neither noise nor undecided by CANDLE. Additionally, “relaxed” recall (RR) values are given that also accept overlapping labels like “C-F” for “C” and vice versa as correct. The best scores for each category and line are highlighted.

Class	Size	CANDLE				kNN				Random Forest			
		Noise	Undec.	R*	RR*	R	RR	R*	RR*	R	RR	R*	RR*
C	226	3.10	34.96	62.86	<b>96.43</b>	50.00	60.62	57.86	67.86	<b>77.88</b>	<b>87.61</b>	<b>77.86</b>	91.43
F	517	2.13	29.79	49.72	<b>94.60</b>	81.62	86.27	80.68	85.80	<b>86.07</b>	<b>91.10</b>	<b>86.08</b>	92.05
G	598	1.84	14.05	80.32	98.21	94.15	95.32	96.02	96.82	<b>96.15</b>	<b>97.83</b>	<b>96.42</b>	<b>98.41</b>
C-F	183	0.55	23.50	<b>87.77</b>	99.28	24.59	<b>99.45</b>	32.37	<b>100.00</b>	52.46	98.91	60.43	99.28
F-G	183	0.55	19.13	<b>87.07</b>	99.32	14.75	<b>100.00</b>	17.01	<b>100.00</b>	43.17	<b>100.00</b>	51.02	<b>100.00</b>

the recall and relaxed recall values of both the  $k$ -nearest neighbor and random forest classifier improve on average, albeit only slightly for the random forest classifier. The undecided labels are, hence, assigned to points that are ambiguous to some extent. The ambiguity is further expressed by the low scores of the other classifiers on the undecided labeled instances. Aside from merely improving scores, this also gives an insight into the nature of the data set.

Intuitively, we would expect the selected noise samples to describe extraordinarily good or bad seasons of some players or just players in general that have a very extraordinary play style. As the other classifiers have a higher accuracy on the noise instances than on the rest, these samples should be fairly representative of their position but outstanding in their execution. On inspection of the list of noise labeled samples from the entire data set, we found our intuition supported by, e.g., the following entries: Amongst the players with most noise seasons are players like Jason Kidd, Chris Paul, Dwyane Wade, and Antoine Walker. All these players are considered exceptionally good players with outstanding performances. Jason Kidd for example is known for his Triple-Doubles (achieving double-digit values in at least three game statistics) records, clearly distinguishing him from the rest of the league. Among the players with only one noise labeled season are players like Boris Diaw ('06) and Eddy Curry ('07) both having the season of their career with outstanding point and rebound counts, or less fortunate players like Hasheem Thabeet ('12) that performed subpar. The list even shares many entries with the MVP awards, though not perfectly. Derrick Rose's season '11 appears amongst the noise samples as does Dirk Nowitzki's season '07 but Steve Nash's seasons '11 and '12 were seemingly more outstanding than his awarded seasons '05 and '06. LeBron James does not appear in the list of noise samples, although his awarded seasons '09 and '12 are very close to being considered noise (plausibility score below 0.1). The “unexpected” performances of outstanding players are, hence, semantically captured by the noise labels.

### 4.3. MNIST-PCA

The MNIST data set is a popular collection of  $28 \times 28$  pixel grayscale images of handwritten digits. Instead of using these 784 dimensions, however, we applied a PCA dimension reduction to 50 features so we can work with much smaller covariance matrices (this is a common approach, also used, e.g., for t-SNE [4]). This does not discard too much information from the data set as the images have, e.g., large areas that are almost always 0 and in general many correlated pixels. The total explained variance of this reduction came up to be 82.55%. As this data set is much larger

**Table 4:** Accuracy (A) for all, non-noise/undecided-labeled, noise-labeled, and undecided-labeled samples of the MNIST-PCA test set. The best accuracy for each row is highlighted.

Class	CANDLE			kNN	Random Forest	RBF-SVM	Log. Regression
	Noise	Undec.	A	A	A	A	A
All	8.40	1.05	(89.90)	97.56	95.46	<b>98.52</b>	90.84
Non-Noise/Undec.	-	-	99.28	98.81	97.46	<b>99.47</b>	93.47
Noise	100.00	-	-	85.62	75.40	<b>88.94</b>	63.57
Undec.	-	100.00	-	85.03	83.67	<b>93.20</b>	82.31

**Table 5:** Recall (R) values for individual classes on the MNIST-PCA test set. Columns marked with \* are computed on all instances that are neither labeled noise nor undecided by CANDLE. The best scores for each category and line are highlighted.

Class	Size	CANDLE			kNN		Random Forest		RBF-SVM		Log. Regression	
		Noise	Undec.	R*	R	R*	R	R*	R	R*	R	R*
0	1380	10.22	0.00	<b>100.00</b>	99.42	99.84	97.97	99.19	<b>99.49</b>	99.92	96.81	98.39
1	1575	2.41	4.76	97.26	99.24	<b>99.66</b>	98.67	99.45	<b>99.43</b>	99.59	97.08	98.15
2	1397	11.02	0.07	<b>99.68</b>	97.21	98.47	94.63	96.78	<b>98.71</b>	99.44	87.83	90.18
3	1428	10.50	0.63	<b>99.45</b>	96.29	98.27	93.98	97.08	<b>97.90</b>	<b>99.45</b>	87.25	91.41
4	1364	7.99	0.37	<b>99.76</b>	97.07	98.00	95.45	96.56	<b>98.39</b>	99.04	92.23	94.08
5	1262	10.30	1.03	<b>99.55</b>	97.07	99.11	94.53	97.68	<b>98.26</b>	<b>99.55</b>	84.79	89.01
6	1375	10.55	0.44	99.51	98.33	99.59	97.24	98.86	<b>98.69</b>	<b>99.75</b>	93.60	96.57
7	1458	5.28	1.17	99.05	98.29	98.97	96.57	97.87	<b>98.77</b>	<b>99.63</b>	92.66	94.87
8	1364	8.21	0.15	<b>99.84</b>	95.67	97.44	92.52	95.12	<b>97.73</b>	99.12	85.56	87.92
9	1391	8.55	1.37	99.12	96.69	98.64	92.52	95.69	<b>97.70</b>	<b>99.20</b>	89.22	92.82

than the others, we did not perform a grid search on the CANDLE model parameters but rather chose them by hand. For  $n$  we used a neighborhood size that gave stable ABID estimates [19], as the approach is related in concept. The value for  $k$  was chosen equal to the tuned parameter of the  $k$ -nearest neighbor classifier. Both  $c$  and  $b$  were chosen intuitively to not exclude too many instances, yet, constraint the model enough to be sure about classifications. Table 4 and Table 5 display our results on this dimensionally reduced MNIST data set, where filtered results are displayed analog to the NBA results. The CANDLE model ( $n = 150, k = 8, c = 3, b = 0.15$ ) labeled a total of 8.4% of test instances as noise and achieved a very good test accuracy only matched by an RBF-SVM model. Without filtering out the noise and undecided instances, however, the RBF-SVM only achieves an accuracy below the CANDLE model. Again, filtering with noise and undecided labels also improved the results of the reference classifiers as all classifiers are less accurate on both categories. Hence, if not deciding a class for a certain portion of instances is acceptable, the CANDLE model outperforms the reference models. In use cases, where, e.g., end-user interaction can be requested to rewrite problematic digits, or an expert can be requested to label problematic instances, being more accurate in “clear cases” might be more important than being more accurate in all cases.

The CANDLE model can also be used to find “weaknesses” of the training and test data. Figure 3 displays some of the test instances labeled as noise by CANDLE. These instances are only some of the more eccentric examples of digits in the MNIST data set and do not represent the essence of their specific digits. Without the additional information of these symbols all being digits, a human observer might not even successfully interpret these examples. While it



**Figure 3:** Some of the MNIST instances that were labeled noise by CANDLE. The digits should be 8, 4, 6, 7, and 2 from left to right but are very untypical for these digits.

is generally attempted to have classifiers generalize so well as to even accept these instances, it comes at the cost of unpredicted behavior when an out-of-domain instance is presented. This problem can be avoided by integrating noise detection into classification. Similar problems can be deduced from the percentages of undecided instances per class. For example, many 1-instances in the MNIST data set have high similarity to 2-, 7-, 8- and 9-instances. Using overlap information can potentially also be used to weigh training instances to better shape the decision boundaries or to clean the training data for the benefit of accuracy on more typical instances.

## 5. Conclusion

In this paper, we introduced a new classification approach based on the assumption that each class approximately describes a manifold. We use this common assumption to obtain a local Mahalanobis distance-based score for how likely a given point lies on the same manifold as given training data. Using these scores we can assign labels to query points as being either noise, undecided, or part of one of the classes. Our experimental results suggest that our approach gives good classification results for all non-noise and non-undecided query points. Yet, it achieves only inferior results whenever noise- and undecided-labels are not used. The approach is, thus, most appropriately used in contexts, where the most difficult decisions can be decided by experts (or alternative models) but automation tools are required to sieve out the easy decisions. As for future work on this subject, an index structure for similar covariance matrices, making the plausibility score  $L$  locally sensitive rather than using global mean and standard deviations, and adding a better selection scheme to reduce the training set would be useful additions. Also, the approach is readily adaptable to other measures like outlier scores which opens up a whole branch of classification models.

## References

- [1] L. Guan, R. Tibshirani, Prediction and outlier detection in classification problems, 2019. [arXiv:1905.04396](https://arxiv.org/abs/1905.04396).
- [2] Y.-C. Chen, C. R. Genovese, L. Wasserman, Density level sets: Asymptotics, inference, and visualization, *J. American Statistical Association* 112 (2017) 1684–1696. doi:10.1080/01621459.2016.1228536.
- [3] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 5500 (2000) 2323–6. doi:10.1126/science.290.5500.2323.

- [4] L. V. D. Maaten, G. E. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research* 9 (2008) 2579–2605.
- [5] S. K. Sen, M. Foskey, J. S. Marron, M. A. Styner, Support vector machine for data on manifolds: An application to image analysis, in: *IEEE Int. Symp. Biomedical Imaging: From Nano to Macro*, 2008, pp. 1195–1198. doi:10.1109/ISBI.2008.4541216.
- [6] D. Li, D. B. Dunson, Classification via local manifold approximation, 2019. arXiv:1903.00985.
- [7] H. Kriegel, P. Kröger, E. Schubert, A. Zimek, Outlier detection in arbitrarily oriented subspaces, in: *IEEE Int. Conf. on Data Mining, ICDM*, 2012, pp. 379–388. doi:10.1109/ICDM.2012.21.
- [8] H. Kriegel, M. Schubert, A. Zimek, Angle-based outlier detection in high-dimensional data, in: *ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2008, pp. 444–452. doi:10.1145/1401890.1401946.
- [9] C. Böhm, K. Kailing, P. Kröger, A. Zimek, Computing clusters of correlation connected objects, in: *ACM SIGMOD Int. Conf. Management of Data*, 2004, pp. 455–466. doi:10.1145/1007568.1007620.
- [10] E. Achtert, C. Böhm, H. Kriegel, P. Kröger, A. Zimek, Robust, complete, and efficient correlation clustering, in: *SIAM Int. Conf. Data Mining, SDM*, 2007, pp. 413–418. doi:10.1137/1.9781611972771.37.
- [11] R. M. Haralick, R. Harpaz, Linear manifold clustering in high dimensional spaces by stochastic search, *Pattern Recognit.* 40 (2007) 2672–2684. doi:10.1016/j.patcog.2007.01.020.
- [12] E. Achtert, C. Böhm, H. Kriegel, P. Kröger, A. Zimek, On exploring complex relationships of correlation clusters, in: *Int. Conf. Scientific and Statistical Database Management, SSDBM*, 2007, p. 7. doi:10.1109/SSDBM.2007.21.
- [13] E. Achtert, C. Böhm, J. David, P. Kröger, A. Zimek, Robust clustering in arbitrarily oriented subspaces, in: *SIAM Int. Conf. Data Mining, SDM*, 2008, pp. 763–774. doi:10.1137/1.9781611972788.69.
- [14] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, M. Nett, Estimating local intrinsic dimensionality, in: *ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2015. doi:10.1145/2783258.2783405.
- [15] M. E. Houle, H. Kashima, M. Nett, Generalized expansion dimension, in: *ICDM Workshops*, 2012. doi:10.1109/ICDMW.2012.94.
- [16] O. Chelly, M. E. Houle, K. Kawarabayashi, Enhanced Estimation of Local Intrinsic Dimensionality Using Auxiliary Distances, Technical Report NII-2016-007E, National Institute of Informatics, 2016.
- [17] L. Amsaleg, O. Chelly, M. E. Houle, K. Kawarabayashi, M. Radovanovic, W. Treeratanajaru, Intrinsic dimensionality estimation within tight localities, in: *SIAM Int. Conf. Data Mining, SDM*, 2019. doi:10.1137/1.9781611975673.21.
- [18] V. Erba, M. Gherardi, P. Rotondo, Intrinsic dimension estimation for locally undersampled data, *Scientific Reports* 9 (2019).
- [19] E. Thordsen, E. Schubert, ABID: angle based intrinsic dimensionality, in: *Int. Conf. Similarity Search and Applications, SISAP*, 2020. doi:10.1007/978-3-030-60936-8\_17.

- [20] M. Aumüller, M. Ceccarelo, The role of local intrinsic dimensionality in benchmarking nearest neighbor search, in: Int. Conf. Similarity Search and Applications, SISAP, 2019. doi:10.1007/978-3-030-32047-8\_11.
- [21] R. Becker, I. Hafnaoui, M. E. Houle, P. Li, A. Zimek, Subspace determination through local intrinsic dimensional decomposition, in: Int. Conf. Similarity Search and Applications, SISAP, 2019. doi:10.1007/978-3-030-32047-8\_25.
- [22] M. E. Houle, E. Schubert, A. Zimek, On the correlation between local intrinsic dimensionality and outlierness, in: Int. Conf. Similarity Search and Applications, SISAP, 2018. doi:10.1007/978-3-030-02224-2\_14.
- [23] E. Thordsen, E. Schubert, MESS: Manifold embedding motivated super sampling, 2021. arXiv:2107.06566.