Strangeness Based Feature Selection for Part Based Recognition

Fayin Li and Jana Košecká and Harry Wechsler
George Mason Univerity
4400 University Dr. Fairfax, VA 22030 USA

Abstract

Motivated by recent approaches to object recognition, where objects are represented in terms of parts, we propose a new algorithm for selecting discriminative features based on strangeness measure*. We will show that* k-nearest neighbour strangeness *can be used to measure the uncertainty of individual features with respect to the class labels and forms piecewise constant decision boundary. We study its properties and generalization capability by comparing it with optimal decision boundary and boundary obtained by k-nearest-neighbor methods. The proposed feature selection algorithm is tested both in simulation and real experiments, demonstrating that meaningful discriminative local features are selected despite the presence of large numbers of distractors. In the second stage we demonstrate how to integrate the local evidence provided by the selected features in the boosting framework in order to obtain the final strong classifier. The performance of the feature selection algorithm and the classifier is evaluated on the Caltech five object category database, achieving superior results in comparison with existing approaches at lower computational cost.*

1. Introduction

In many supervised learning tasks, the input data is often represented by a large number of often high dimensional features. Even state-of-art learning algorithms cannot overcome the presence of a large number of weakly relevant or irrelevant features. Once a good set of features is obtained, even the very basic and simple classifiers can achieve high performance. Additional benefits of feature selection are in reducing the measurement and storage requirements, reducing the complexity of learned models, defying the curse of dimensionality to improve prediction performance and facilitating data visualization and data understanding.

In general setting, given the training features $F = (F_1, \cdots, F_N) \in \Re^{d \times N}$, where F_i is a point in $\mathbb{R}^d$, there are two different feature selection directions: one is to select the optimal subspace along the column direction of the feature matrix F - *variable selection*; the other one is to select the optimal sub-instance along the row direction of F - *feature instance selection*. The first direction is widely researched in the machine learning field, where one assumes that each instance of F has some contribution for classification and tries to find the optimal subspace and compact representation. The second direction is commonly encountered in the computer vision community, in the context of part based representations of objects and object categories.

Our work is motivated by several recent approaches to weakly supervised learning of object categories as well as general object recognition, which consider representations of objects in terms of parts [4] . Learning of the object parts for different categories which constitute visual vocabularies used to built object models is often the first stage of existing approaches. Most frequently this stage is addressed by clustering local features corresponding to salient regions in the image [2]. The number of detected features is typically quite large, with many features coming from the background, yielding large visual vocabularies, with many superfluous clusters. Furthermore, k-means clustering is often unstable when the space is populated by a large number of distractors. In other recognition tasks, such as recognition of object instances, actual instances of discriminative features need to be learned to obtain good models [9]. This stage can hence greatly benefit from the *feature instance selection* process.

To tackle these issues we propose a new feature selection algorithm based on *k-nearest-neighbour strangeness* measure; k-NN strangeness is the ratio of the sum of k nearest distances from the same class divided by the sum of k nearest distances from all other classes. We first study strangeness properties and show how they can be used to measure the uncertainty of the individual features with respect to the class labels and to construct the decision boundary. We then introduce the proposed feature instance selection algorithm and

test it both in simulation and real experiments. We demonstrate the performance of the feature selection algorithm on a object category recognition task demonstrating that meaningful discriminative local features are selected despite the presence of a large number of distractors. The selected features constitute different instances of parts. In the second stage we show to integrate the local evidence provided by parts in the boosting framework, with the strangeness used as weak hypothesis. The second stage can be viewed as another feature selection strategy, in which boosting will select the most discriminative parts.

2. Related work

Feature Selection Different feature selection algorithms can be broadly divided into two categories: *filters* and *wrappers*. The filter approaches evaluate the relevance of each feature (subset) using the data set alone, regardless of the subsequent learning phase. RELIEF method [12] and information theoretic methods [15, 9] are the representatives of this class. The philosophy behind the information theoretic methods is that the mutual information between relevant features and class labels should be high. In computer vision an example of this approach is [2], where scale-invariant image features are extracted and ranked by a likelihood or mutual information criterion. On the other hand, the wrapper approaches [6] use a learning algorithm to evaluate the quality of each feature (subset). In the learning phase Boosting [16, 14], Bayesian approach [4], decision trees [9] were used in the past and the feature relevance was assessed by the estimation of the classification accuracy. Wrappers are usually more computationally demanding, but can be superior in accuracy when compared with filters. Both approaches involve combinatorial search through the space of possible feature subsets with different types of heuristics.

Strangeness The strangeness measure used in our approach is the ratio of the sum of the k nearest distances from the same class to the sum of the k nearest distances from all other classes. The approach falls into the category of non-parametric data driven approaches for classification, such as prototype and nearest neighbour methods. In case of parametric approaches, Bayesian inference is often used to estimate the posterior probability of the class. However, the optimality of the Bayesian method is based on the assumption that the data we observe are *generated* according to one of the distribution models in *the chosen class of models*. While this assumption is attractive for theory, it rarely holds in practice. In the context of general classification tasks, instead of assuming a family of models, Vovk et al [13] introduce an *individual strangeness measure* and construct the confidence machine using the algorithmic theory of randomness and transductive inference. While in inductive inference, where training data is used to find some approximation of functional dependency between data and class labels (which is then evaluated at points of interest), in transductive inference the value of the function is evaluated only at points of interest. The simplest method of this type of inference is k-nearest neighbour method. The strangeness α_i of a particular example $\mathbf{x}_i$ measures the uncertainty of that example with respect to its label and all other examples: the higher the measure, the higher the uncertainty. It is, in fact, the discriminanation ability of that example. Hence, strangeness measure can be used either for classification or feature selection, which will be shown in the later sections.

3. Strangeness Measure

Several strangeness definitions were proposed [5, 8], which need complex learning strategies and high computational costs. There are several simpler definitions which do not require complex learning procedures. If the example of class j is sampled from a Gaussian model, the distance from example $\mathbf{x}_i^j$ to the mean $\bar{\mathbf{x}}^j$ is defined as the strangeness:

$$\alpha_i = \|\mathbf{x}_i^j - \bar{\mathbf{x}}^j\|, \text{ where } \bar{\mathbf{x}}^j = \frac{1}{N_j}\sum_k \mathbf{x}_k^j.$$

Without any assumption about distribution D of $z = (\mathbf{x}, y)$, where y is the class label, k-nearest neighbor classifier is widely used in [10, 13] to define the strangeness measure if the examples are measurable in some metric space. Assume we have C classes, for class $c = 1, \cdots, C$, let us denote the sorted sequence (in ascending order) of the distances of example $\mathbf{x}_j^c$ from the other examples with the same classification c as d_j^c and d_{jl}^c will stand for the l^{th} shortest distance in this sequence. Let d_j^{-c} denote the sorted sequence of distances containing examples with a classification different from c. For each example, the individual strangeness measure is assigned as:

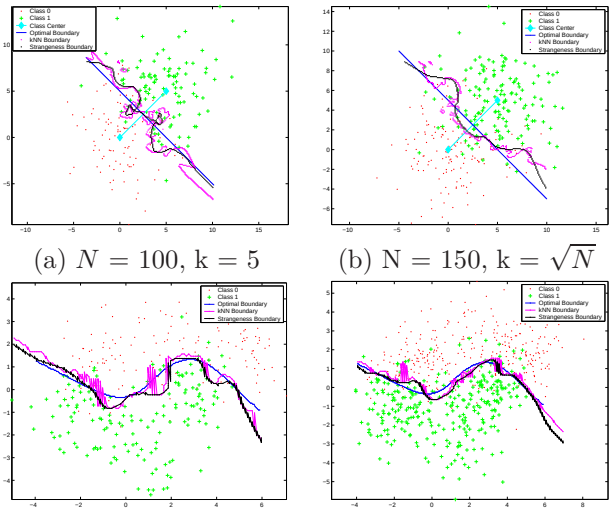
$$\alpha_j = \frac{\sum_{l=1}^k d_{jl}^c}{\sum_{l=1}^k d_{jl}^{-c}}. \quad (1)$$

The measurement for strangeness is the ratio of the sum of the k nearest distances from the same class to the sum of the k nearest distances from all other classes. This definition of strangeness is very natural and straightforward. An example is considered strange

if it is in the middle of examples labeled in a different way and is far from the examples labeled in the same way. The strangeness of an example increases when the distance from the example of the same class becomes bigger or when the distance from the other classes becomes smaller. The strangeness defined in Equation 1 is related to k-nearest neighbor classifier (k-NN). However, for multi-class classification, the definition in Equation 1 does not consider the frequency of each class in the neighborhood of the example, as does in k-NN classifier. As the result, we modify the definition in Equation 1 and re-define the k-NN strangeness as:

$$\alpha_j = \frac{\sum_{l=1}^{k} d_{jl}^{c}}{\min_{n,n\neq c} \sum_{l=1}^{k} d_{jl}^{n}}, \qquad (2)$$

where in the denominator is the class, with the minimal sum of k-NN distances. In the following subsection, we will discuss its properties and show how it is related optimal decision boundary and the posterior $P(c_i|\mathbf{x}_i)$.



(a) $N = 100$, k = 5 (b) N = 150, k = $\sqrt{N}$

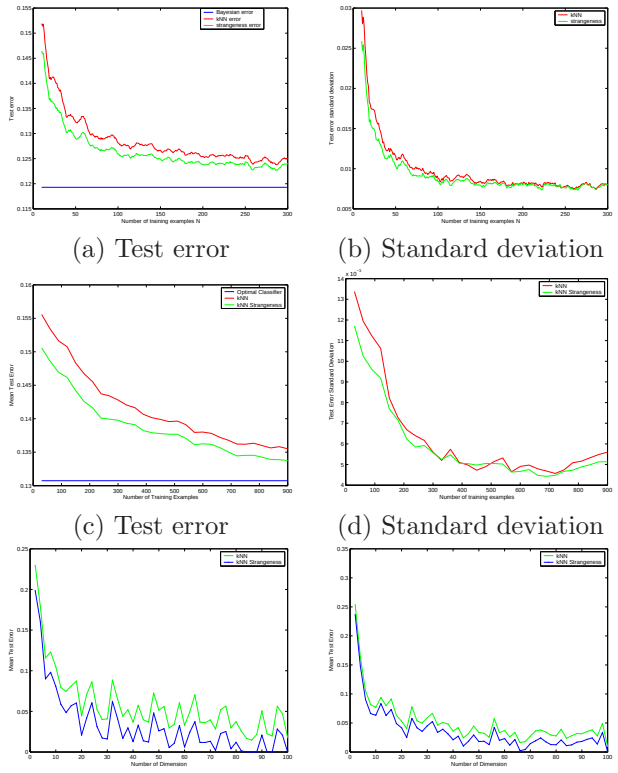
(c) $N = 150$, k = $\sqrt{N}$ (d) N = 300, k = $\sqrt{N}$

Figure 1. Top: the boundaries with different k and different N for two Gaussians. Bottom: the boundaries constructed with different N for Gaussian mixtures.

3.1. k-Nearest Neighbor Strangeness

In this section we will study the properties of k-nearest neighbor strangeness (as defined in Equation 2), how it can be used to build the decision boundary between classes and how it is related to the discrimination ability of each example. The Cover-Hart theorem proves that asymptotically the generalization error of 1-NN classifier can exceed by at most twice the generalization error of the Bayes optimal classification rule. They also showed that the k-NN error approaches the Bayes error (with factor 1) if $k = O(\log n)$ [1]. The generalization power of k-NN classifier enables the k-NN strangeness to have the similar properties. On average, the examples with $\alpha = const$ build the piecewise linear boundary between class c_i and all other classes. Asymptotically, the examples with $\alpha = 1$ will build the optimal boundary between two classes. Those examples can be considered as samples from the optimal Bayes classification boundary which serves as the ground truth if the data distribution and prior are known. To demonstrate this effect, consider a two-class classification first. Let examples $(z_1, \cdots, z_n) = ((\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be drawn independently from the same distribution over $Z = X^d \times Y$ where Y is the label space $\{0, 1\}$. For each class c_i, the data is generated independently from a Gaussian distributions $P(\mathbf{x}|c_i) = N(\mathbf{x}; \mu_i, \Sigma_i^{-1})$ and priors $p_i = P(c_i)$, $i = 0$ or 1. Let the means of two classes be $[0, 0]^T$ and $[5, 5]^T$ with the same covariance matrix $\Sigma = diag\{\sigma, \sigma\}$. Both classes have the same number N of samples, that is $p_0 = p_1 = 0.5$. In 2D separable case the advantage of the strangeness measure is not so apparent. In real applications the classes are rarely well-separable; and the data are often in a high dimension space. Therefore, we focus on the comparison with non-separable data sets in a high dimensional space. Fig. 1 a) and b) shows two Gaussians with $\sigma = 3$, and a different number of training examples N with $k = \sqrt{N}$. Fig. 1c) and d) shows the optimal boundary, and the boundaries of strangeness and k-NN, respectively, when two classes are mixtures of Gaussian distribution. Class 0 has the three modes with the means $\{[2, 2], [-1, 1], [5, 2]\}$ and covariance matrices $\{\text{diag}([1.5, 1.5]), \text{diag}([1, 1]), \text{diag}([1, 1])\}$, respectively. Class 1 has also three modes with the means $\{[1, -2], [-2, -1], [3, 0]\}$ and the same covariance matrices as class 0. Each mode has the same weight in both classes. For each class, N training examples are randomly drawn. Note that while both boundaries are far from the optimal boundary, the boundary constructed by strangeness is much more smooth and closer to the optimal boundary. When k is small, the strangeness smoothes many isolated regions created by k-NN classifier. If we consider the problem in a regularization framework, strangeness introduces a smooth penalty term, which is defined through the examples with the parameter k. The boundaries constructed depend highly on the training examples while they both converge to the optimal boundary as $N \to \infty$. Let's now consider the generalization ability of both classifiers and evaluate their test errors. For each N, different training and testing sets are sampled in 100 trials. Fig. 2 (a)-(d) shows the optimal Bayesian error,

test errors of k-NN and strangeness classifiers, and the corresponding error standard deviation over the trials. We evaluate these for 2D Gaussian distributions (a),(b) and mixtures of Gaussians (c),(d). For 2 (e) and (f) considers two d-dimensional Gaussian distributions with means $[0, \cdots, 0]$ and $[5, \cdots, 5]$, respectively. Assume they have different covariance matrices such that the optimal classification boundary is no longer a hyperplane. The two covariance matrices are randomly generated. In d dimensional space, each class has N training examples randomly sampled from the distributions above. Another 10000 examples are randomly generated for testing. Fig. 2e), f) show the test errors of k-NN and strangeness ($\alpha = 1$) in different dimensional spaces, d from 2 to 100. It clearly shows the strangeness has better performance over k nearest neighbor classifier no matter the dimensionality of the representation. Note that the error of strangeness and its standard deviation are always lower than those of the corresponding k-NN classifier, which is consistent with the conclusion from the comparison of the boundaries. The smooth term of the classification function



(a) Test error (b) Standard deviation

(c) Test error (d) Standard deviation

(e) Test error $N = 100$ (f) Test error $N = 200$

Figure 2. The test errors and their standard deviations. Top: two Gaussians. Middle: Gaussian mixtures. Bottom: high dimensional Gaussians; the test errors with respect to the dimensionality given fixed N of training examples.

reduces the test error and hence improves the generalization capability of the algorithm.

So far we have only compared the classification performance of the strangeness measure. Note however, that although both classifiers have close performance, the strangeness not only yields the "bare prediction" as k-nearest-neighbor does, it also gives the "confidence" or "reliability" of the prediction: the higher the measure, the higher the uncertainty of the prediction. It can be further shown that the strangeness measure has a monotonic relationship with margin, posterior and odds. This is the key property which we will use for feature selection and the classifier design.

4. Feature Instance Selection Algorithm

In the previous section we presented the definition of the strangeness, studied its properties and its generalization capability. Next we will show how strangeness can be used to evaluate the feature relevance.

In order to deal with the large variation of object appearance, due to occlusions, pose variation, deformation, and size, many appearance-based approaches to object recognition characterize the objects by image features, corresponding to local image regions. These can be either directly image patches [9] or affine invariant regions and their associated descriptors [4]. Each image is represented by M_i features $\{g_j\}$ in d dimensional space. Many of the generative approaches mentioned earlier [2, 4] use k-means clustering in the first stage to create a visual vocabulary of parts. The number of clusters and clustering algorithms can have a great influence on the performance and generalization ability of the final classifier. Since the features from the background are assumed to be distributed uniformly in the descriptor space, a large number of irrelevant features may yield large number of clusters and overwhelm the relevant features in the clustering algorithm. As we will demonstrate next, a simple and efficient algorithm for discarding the irrelevant features and selecting the discriminative features for later learning stages can successfully tackle some of the above mentioned problems. The algorithm is based on the strangeness measure α which is used to evaluate the relevance between each local feature and the class label of whole image.

Strangeness Feature Instance Selection Algorithm 1 is an iterative backward elimination method. The algorithm repeatedly iterates over the feature set and updates the set of chosen features. There is one threshold in the algorithm γ, which determines the features to be eliminated in each iteration and controls the largest strangeness, that is, the minimal margin, of the chosen features in the end. The algorithm can be applied very efficiently if suitable data structures are used, because

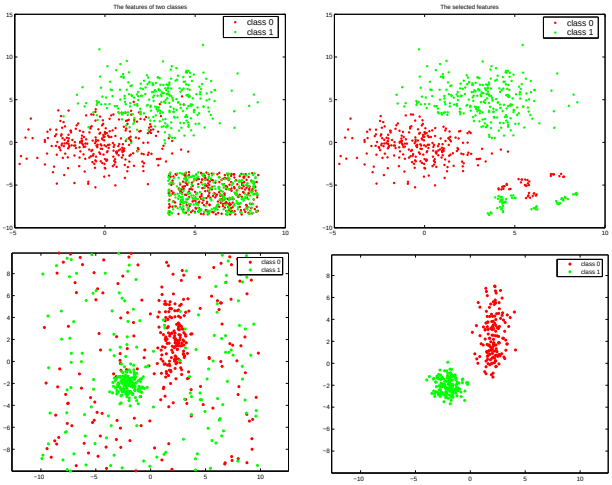
Algorithm 1 Strangeness Feature Instance Selection

1. Given local features $\{g_i\}$ in $\mathbb{R}^d$ and class label.
2. Compute the strangeness of each feature g_i based on Equation 2.
3. Initialize the threshold of strangeness γ.
4. **for** t = 1,2,...,T
   - Select the features $\{g_k\}$ with the strangeness $\alpha_k \geq \gamma$.
   - Discard $\{g_k\}$ and update the strangeness of remaining features.
   - If the strangeness of all features is less than γ, terminate.
5. end **for**

only small portion of strangeness values needs updating in each iteration. Compared with other feature selection algorithms, Algorithm 4 not only has the advantage of *filter* approaches – evaluating the relevance of feature and simple, but also have the properties of *wrapper* approaches – related to the predictor generalization performance.

5. Experiments and Evaluation

In this section, we demonstrate the behavior and performance of the Strangeness Instance Feature Selection Algorithm on a small synthetic two-class classification problem. Consider two classes with different kinds of features sampled from different distributions. As shown in Fig. 3(a), the first class has two kinds of features sampled from two distributions: Gaussian distribution D_1 with mean $[0,0]^T$ and standard deviation $\sigma = 2$, and uniform distribution D_0 over region $(3.5, 8.5) \times (-8.5, -3.5)$; the second class also has two kinds of features sampled from two distributions: Gaussian distribution D_2 with mean $[3,5]^T$ and standard deviation $\sigma = 2$, and uniform distribution D_0 over region $(3.5, 8.5) \times (-8.5, -3.5)$. For each distribution in each class, 300 points are randomly sampled as the training data set. Note that two different classes have the features sampled from the same distribution D_0, which, in the context weakly supervised object recognition task, would correspond to background features. Fig. 3(b) shows the selected features. As we can see from the figures, the most informative feature points are kept and most features with low discriminative ability are discarded. Only a very small number of features is chosen from D_0. This demonstrates that the proposed feature selection method effectively discards irrelevant features and hence, can precede many of the standard learning algorithms which attempt to learn generative models. Note that in the second example



(a) (b)

Figure 3. The features in the original data sets and the results after feature selection.

all the disctractor features have been successfully eliminated. This number of remaining features is function of threshold γ.

The presented feature selection algorithm is next applied on weakly supervised object category recognition using Caltech database. The more detail information about the data base can be found in [3]. Fig. 4 shows the original features detected and the features chosen by the algorithm. As expected, most of the selected features are on the objects while most background features are discarded. After the initial feature selection,

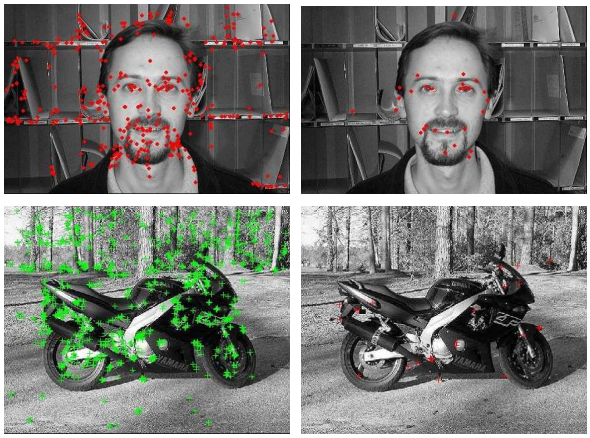
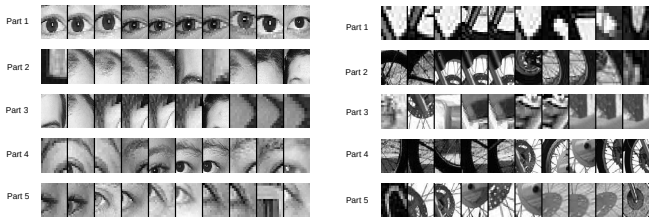


Figure 4. The original features detected and the selected feature set.

most local features in training images now have strong relevance with respect to the classification and the complexity of the classification task is highly reduced.

5.1. Final classifier

In this section, we show how selected local features can be used as local classification evidence which can be integrated in the boosting framework with the strangeness based weak classifier. After feature selection, each training I_k image is represented by the selected feature set $\{g_j^k\}$, with each feature having its associated strangeness computed from Equation 2. Considering strangeness as the base classifier, we can apply the AdaBoost algorithm on the selected feature set directly. However, several features may be extracted from almost the same location of the same object yielding redundant information. If each feature is considered as a weak classifier as in [11], the final strong classifier will be overfitting and have the low generalization capability. For example, eye is a very important feature to distinguish face from other objects. If the final strong classifier has several "eye" weak learners, it has high probability of misclassifying the test face if the "eye" feature is not detected in the image. In order to achieve high generalization ability of the final classifier, we first reduce the information redundancy among local features by clustering them into parts, and then model the local classification evidence by a model-free, non-parametric approach using the strangeness of each feature instance in each part. Figure 5 shows the parts of the motorbike and faces categories after feature selection based on strangeness and clustering. The evidence provided by individual parts is then integrated in the second stage in the boosting framework, where we design a strangeness based weak learner for each part. This can be viewed as another feature selection stage, in which boosting will select the most discriminant and reliable parts.



(a) Face Parts (b) Motorcycle Parts

Figure 5. Grouped object parts - weak rules in boosting.

Starting with the training data set, we have now each object category c represented by P parts, each of which has N_i feature instances $G_i^c = \{g_j^i\}_{j=1}^{N_i}$. Instead of parametric modelling of the clusters, we keep their feature instances as the training gallery, apply the base classifier on P parts and learn the coefficients and thresholds of weak learner through validation data set. Given the validation image V_i and its local features descriptor $\{g(V_i)_j\}$ with putative object label c, the matched features $\{\tilde{g}(V_i)_j^c\}_{j=1}^P$ are found which are the closest feature from $\{g(V_i)_j\}$ to each part of class c in the gallery. Then the strangeness $\{\alpha_i^c\}$ of $\{\tilde{g}(V_i)_j\}_{j=1}^P$ are computed with the assumption of putative class c. With C classes in the training gallery, C groups of strangeness are obtained for each validation image. If M validation images are given for each class, for each part of each class, we have M positive strangeness measures and $M(C-1)$ negative ones. Our weak hypothesis is to select the matched feature $\{\tilde{g}(V_i)_j\}_{j=1}^P$ and the strangeness threshold T_j for each part of the class. The Algorithm 2 describes the strangeness based weak learner.

In this manner we can obtain a weak classifier for each of the P parts, where the thresholds and the coefficients of the weak classifiers are learned in the validation stage. The Strangeness Weak Learner is model-free and non-parametric, and as simple as the stump function. The main computational burden is the calculation of strangeness of $g(V_i)_j$ with putative label c, since it needs distance from $g(V_i)_j$ to all features in the training gallery. However, such computation can be done prior to Boosting and weak learner finder. The remaining calculations in Boosting are very inexpensive. Drawing an analogy between weak classifiers and features, this learning model is another aggressive feature selection mechanism for selecting a small set of "good" features which nevertheless have significant variety. Finally C group of coefficients $\{\beta_t^c\}_{t=1}^P$ are obtained, which tell the importance of each part for each subject. The coefficients are then normalized such that $\sum_{t=1}^P \beta_t^c = 1$. Final decision rule for the query image Q then has a following form:

$$f^c(Q) = \beta_1 h_1(Q) + \ldots + h_p(Q).$$

The testing proceeds in the way similar to the validation stage. We demonstrate the performance of the feature selection algorithm on the object category recognition tasks using 4 object categories: motorbike, airplane, faces and cars(side) and background class. Instead of just discriminating the object category from background as in [4, 3, 11, 2], we propose a two-stage hierarchical boosting learning to distinguish the object from both the background and other objects. At first, the strangeness feature instance selection algorithm is applied between objects and background examples and a two-class boosting learner is learned to distinguish all

Algorithm 2 Strangeness Weak Learner

- ***Input***: Training gallery $\{G_j^c\}_{j=1}^P, c = 1, \cdots, C$, where G_j^c is the feature instance set of jth part of class c, and validation images $\{V_i, i = 1, \cdots, MC\}$ and associated feature $\{g(V_i)_k\}$.
- ***Strangeness computation***: For each part j of class c, find the nearest feature $\tilde{g}(V_i)_j$ between $\{g(V_i)_k\}$ and G_j^c. The strangeness of $\tilde{g}(V_i)_j$ is then computed as defined in equation 2 with the putative the class label c of V_i. Each part of class c now has MC strangeness $\{\alpha_k^c\}_{k=1}^{MC}$, M of which are positive and $M(C-1)$ are negative.
- ***Strangeness sorting***: For each part j of class c, let $\pi(1), \cdots, \pi(MC)$ be the permutation such that

$$\alpha_{\pi(1)}^c \leq \alpha_{\pi(2)}^c \leq \cdots \leq \alpha_{\pi(MC)}^c.$$

- ***Select the threshold of weak learner***: For each part j of class c, find the best position s such that the maximal classification rate is achieved:

$$rate(j) = \max_s \sum_{k=1}^{s} w_{\pi(k)} h(\alpha_{\pi(k)})$$

where $h(\alpha_{\pi(k)})$ is 1 if $\alpha_{\pi(k)}$ is positive and 0 otherwise. Then the threshold of current weak learner is:

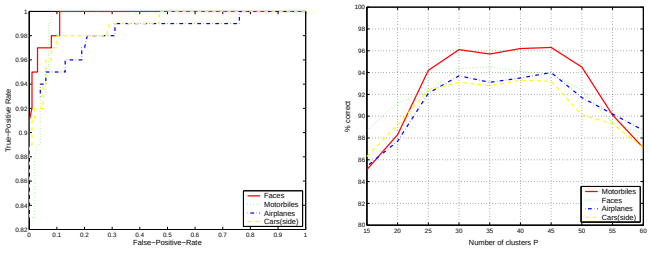
$$\theta(j) = \frac{\alpha_{\pi(s)} + \alpha_{\pi(s+1)}}{2}.$$

- ***Select best weak learner***: Find the best part $m = \max_j rate(j)$. Then the best weak learner of current round is the mth part with the best threshold $T_m = \theta(m)$. Update the weight w_k and compute coefficient β_t according to error $1 - rate(m)$.

object categories from the background category. Based on the features selected in the first stage, further feature selection is done and another one-vs-all boosting learner is used to classify different object categories. In the second stage, the label of query image Q is predicted by $\arg\max P(c|Q)$. It is necessary to use these two stages. Since the background features are uniformly distributed in feature descriptor space, we cannot model the background with parts. As a result we cannot reliably estimate $P(background|Q)$. Given the estimated $P(c|Q)$, it is very hard and almost impossible to find a threshold τ such that Q is background if $\max P(c|Q) \leq \tau$. To avoid estimating such a threshold, the boosting effectively deals with the background. The high performance can be achieved in this stage since more features will be used, some of which have no discriminative power for classification of object categories but have the ability to distinguish objects from background.

For each object, we randomly sample 30 images as the training gallery, 30 images as the validation data set and use the remaining images and the background images for testing. The features are detected by affine covariant regions and represented by SIFT descriptor.

Figure 6(a) shows the ROC curves of our approach on the first database when $P = 30$ and $k = 5$. Figure 6(b) shows the performance with respect to the number of parts P. Table 1 shows the equal error rates of our approach compared with the other two approaches. Our method is a little better than both of them except for the faces. From the results in Figure 6(b) we can see our approach is very stable when the number of parts P is in the range $[25, 50]$. When P is too small or too large, the classifier learned performs poor. When P is small, too little evidence is integrated from the local parts and the final strong classifier does not have enough discriminative power. When P is too large, similar features may have multiple clusters and redundant information exists between weak hypotheses; thus the final strong classifier will be overfitting.



(a) (b)

Figure 6. (a) The ROC curve for image classification on the faces, motorbikes, airplanes and cars(side) data set used by Fergus et al. [4]. (b) The equal error rates with respect to the number of clusters P.

Table 1. The ROC equal error rates on the database used by Fergus et al. [4]

Dataset	Our approach	Fergus	Opelt
Motorbikes	96.1%	92.5%	92.2%
Faces	94.4%	96.4%	93.5%
Airplanes	93.7%	90.2%	88.9%
Cars(side)	93.1%	88.5%	83.0%

In the second stage, a one-vs-all boosting classifier

is learned on the features selected in the first stage. It distinguishes each object from all other object categories, not just at the level of chance as shown in Table 2 in [4]. The work in [3, 11, 2] did not report how their approaches perform on the separation of each category from the others. Table 2 presents the performance of our learning approach across the four classes. Very good recognition rates are achieved. The model for each object successfully rejects the input images from other objects.

Table 2. The performance of the final strong classifier in the second stage on the database used in [4]

Dataset	Motorbikes	Faces	Airplanes	Cars
Motorbikes	93.1%	2.5%	1.6%	2.8%
Faces	1.1%	93.4%	4.5%	1.0%
Airplanes	2.0%	0.0%	95.4%	2.6%
Cars(side)	2.1%	0.0%	6.9%	91.0%

6. Conclusions

We have described a new feature instance selection algorithm based on strangeness measure. In simulation, we have demonstrated its properties and relationship to some baseline classifiers. The proposed algorithm was tested on object category recognition tasks assuming representations of objects in terms of parts. We have shown that the algorithm selects meaningful features and achieves better or comparable classification accuracy at a fraction of the computational cost. Although the presented work was largely motivated by the problem of learning of models for object recognition, the outlined algorithm is applicable in general settings. In the future we plan to extend this approach to a variable feature selection and test the accuracy of the final classifiers on the available benchmark datasets. We are also currently pursuing more detailed theoretical analysis of the bounds on error rates, convergence of the proposed algorithm and connections between other related methods [7].

References

[1] T. Cover and P. Hart. Nearest neighbor pattern classifer. *IEEE Trans. on Inforamtion Theory*, 13, 1967.

[2] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV, Nice, France*, 2003.

[3] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV, Nice, France*, 2003.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–271, 2003.

[5] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of 14th Conference on Uncertainty in Artificial Intelligence*, pages 148–155, 1999.

[6] R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence*, 97:273–324, 1997.

[7] D. Koller and M. Sahami. Toward optimal feature selection. In *13th International Conference on Machine Learning*, pages 284–292, 1995.

[8] T. Melluish, C. Saunders, I. Nouretdinov, and V. Vovk. Comparing the bayes and typicalness framework. In *Proceedings of 12th European Conference on Machine Learning*, volume 2167, pages 350–357, 2001.

[9] M. V. Naquet and S. Ullman. Object recognition with informative features and linear. In *ICCV, Nice, France*, 2003.

[10] I. Nouretdinov, T. Melluish, and V. Vovk. Ridge regression confidence machine. In *Proceedings of 18th International Conference on Machine Learning*, pages 385–392, 2001.

[11] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proceedings of European Conference on Computer Vision*, 2004.

[12] M. Robnik-Sikonja and I. Kononenko. Theoretical and exmpirical analysis of ReliefF and RReliefF. *Machine Learning Journal*, 53:23–69, 2003.

[13] C. Saunders, A. Gammerman, and V. Vovk. Tranduction with confidence and credibility. In *Proceedings of International Conference on Artificial Intelligence*, 1999.

[14] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *International Conference on Computer Vision and Pattern Recognition*, 2004.

[15] N. Vasconcelos. Feature selection by maximum marginal diversity: Optimality and implications for visual recognition. In *CVPR, Madison, Wisconsin*, 2003.

[16] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.