Simpler non-parametric methods provide as good or better results to multiple-instance learning.

Ragav Venkatesan, Parag Shridhar Chandakkar and Baoxin Li
Arizona State University, Tempe, AZ, USA
ragav.venkatesan@asu.edu, pchandak@asu.edu, baoxin.li@asu.edu

Abstract

Multiple-instance learning (MIL) is a unique learning problem in which training data labels are available only for collections of objects (called bags) instead of individual objects (called instances). A plethora of approaches have been developed to solve this problem in the past years. Popular methods include the diverse density, MILIS and DD-SVM. While having been widely used, these methods, particularly those in computer vision have attempted fairly sophisticated solutions to solve certain unique and particular configurations of the MIL space.

In this paper, we analyze the MIL feature space using modified versions of traditional non-parametric techniques like the Parzen window and k*-nearest-neighbour, and develop a learning approach employing distances to* k*-nearest neighbours of a point in the feature space. We show that these methods work as well, if not better than most recently published methods on benchmark datasets. We compare and contrast our analysis with the well-established diverse-density approach and its variants in recent literature, using benchmark datasets including the Musk, Andrews' and Corel datasets, along with a diabetic retinopathy pathology diagnosis dataset. Experimental results demonstrate that, while enjoying an intuitive interpretation and supporting fast learning, these method have the potential of delivering improved performance even for complex data arising from real-world applications.*

1. Introduction

Multiple-instance learning (MIL) is a setting where labels are provided only for a collection of instances called bags. There are two types of instances: negative instances, which are found in either negative bags or positive bags, and positive instances, which are found only in positive bags. While a positive bag should contain at least one inherently positive instance, a negative bag must not contain any positive instances. In MIL, labels are not available at the instance level. It is interesting to note however that the label-space is the same for both at the bag level and at the instance level. One may attempt to learn instance-level labels during the training stage, thus reducing the problem to an instance-level supervised classification. Alternatively, one may also localize and prototype the positive instances in the feature space and rely on the proximity to these prototypes for subsequent classification.

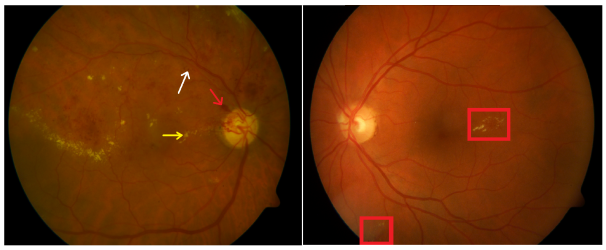

Figure 1. DR image classification as a MIL problem.

MIL is an ideal set-up for many computer vision tasks and examples of its application include object tracking [4], image categorization [9] [26] [28] [12], scene categorization [20] and content-based image retrieval [36]. In particular, MIL can be an especially suitable model for medical image-based pathology classification and lesion detection-localization, where an image is labeled pathological just because of one or a few lesions localized to small portions of the image. Medical images collected in a clinical setting may readily have an image-level label (either normal or various levels of pathology) while lacking the exact location of the lesion(s). Figure 1 illustrates such an example: color fundus images of eyes affected with different pathologies of diabetic retinopathy (DR). It is easy to notice that, although majority of the image looks normal, a small retinal landmark is enough to alter the label of the image from normal to pathological. In a MIL formulation for this problem, each image can be considered a bag and patches of images can be considered instances.

Over the years, many methods have been proposed to solve the MIL problem [10] [29] [8] [2]. The most fun-

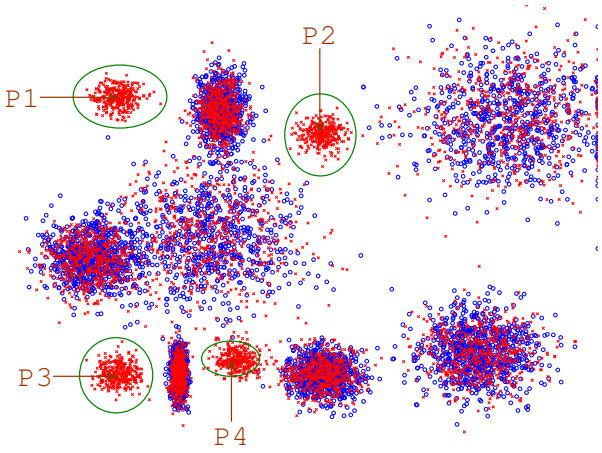

Figure 2. An illustrative feature space for multiple-instance setting. The 'x' in red represents all instances from positive bags and the 'o' in blue represents all instances from negative bags.

damental one is the diverse density approach [19], which has been built upon by many variants [35] [24] [9]. Diverse density is in its basic sense, a function so defined over the feature space such that it is high at any point in the feature space that is close to instances from positive bags while being far away from instances from negative bags and vice-versa. The various local maximas in this function are positive instance prototypes and any instance that is closer to these prototypes are labeled inherently positive instances. Other types of methods also exist in this setting [5] [3] [27] [31].

MIL has many different variants and perspective to its definition and indeed most MIL solutions are application centric [1]. This can be easily seen from table 1. *Earlier methods perform as good or better in the MUSK dataset than the ones published recently although the recent methods perform better on more complex tasks* but for certain exceptions. In this course of research while many particular and complicated solutions are sought after, MIL has never been sufficiently analyzed using traditional non-parametric learning methods. Despite the recent advances, MIL remains a challenging task as the feature space may be arbitrarily complex, the ratio of positive to negative instances can be arbitrarily low in a positive bag, and (by definition) no labeling information is directly available for positive instances.

To illustrate these factors, we simulate a typical MIL feature space as depicted in figure 2. Each instance belonging to a particular cluster is independently drawn from a normal distribution that defines the said cluster. While positive bags can draw a subset of random cardinality of instances from negative distributions, negative bags cannot draw any data from positive distributions. Every positive bag must have at least one instance sampled from a positive distribution (marked in green ellipses $P1$ through $P4$). The centroids of these clusters would be the ideal positive instance prototypes that a MIL algorithm should identify. With the help of this illustration, it is not difficult to imagine that, one or few noisy negative instances coming close to a true positive instance prototype could lower the diverse density drastically and thus lead to a dramatic decrease in performance, and herein lies a core argument to the MIL definition - the strictness of positive neighbourhood. We show that DD-based algorithms are not tolerant even to a single negative instance in an arbitrary positive instance neighbourhood. Such strict assumptions are not suitable for real-world (medical imaging) data wherein the feature space can be noisy.

In this paper, we propose modifications to traditional non-parametric methods adapting them to MIL. We demonstrate their effectiveness against DD taking into consideration the complex arrangements of a typical MIL feature space. In particular, the formulation aims at easing the dramatic impact of noisy negative instances on instance-prototyping in DD-based approaches. The formulation draws intuition from k-nearest-neighbour classification and thus leads readily to an efficient learning algorithm. It employs an aggregated and weighted distance measure computed from any point to its neighbouring instances labeled according to their respective parent bags, conforming to MIL requirement. Analysis with simulated data and experiments with real data in comparison to existing state-of-the-art approaches suggest that the proposed method, while enjoying simplicity in formulation and learning, has the potential of delivering superior performance for challenging benchmark datasets.

The remainder of the paper is organized as follows. Section 2 cites related works, while Section 3 describes the proposed method. Section 4 presents the experimental setup and discusses results on the various evaluation datasets. Section 5 provides concluding remarks.

2. Related Works

MIL was first introduced for the problem of drug activity prediction [10], where axis-parallel hyper-rectangles (APR) were used to design three variants of enclosure algorithms. The APR algorithms tried to surround at least one positive instance from each positive bag while eliminating any negative instances inside it. Any test bag was classified positive as long as it had at least one instance within the APR. Conversely a bag was classified as negative when it had no instance represented within the APR.

The first density-based formulation of MIL was diverse density (DD) [19]. DD is not a conventional density but is rather defined as the intersection of the positive bags against the intersection of the negative bags. It is a measure that is high at any point on the feature space x if x is closer to pos-

itive instances *and* is farther away from negative instances. The local maxima of DD would yield a potential concept for the positive instances. Several local maxima can yield several prototypes of positive instances that can be far apart in the feature space. Some of these prototypes can be separated by other negative instances. The concept point of a diverse density in a MIL feature space was defined as,

$$\arg\max_{x} \prod_{i} Pr(x = t|B_i^+) \prod_{i} Pr(x = t|B_i^-). \quad (1)$$

These local maxima were termed as instance prototypes. A noisy-or model was used to intuitively maximize the DD in Equation 1. This was further developed to assume more complicated and disjoint concepts in EMDD and further developed by other methods including DD-SVM and Accio [35] [9] [24]. The major drawback of the diverse density arises in a situation where the distribution of negative instances is noisy. In other terms, if one instance prototype has a negative distribution closer to the prototype than the others, then its diverse density is largely lower than that of the others, as DD unfairly favours the distribution of positive instances that is farther away from negative instances than those that are relatively closer. This makes it hard to define that particular prototype in such situations. Even the presence of one noisy negative instance near the potential instance prototype can lower the DD drastically as we show in the later sections. In figure 2 the prototype $P4$ was the twenty second largest local maxima in the DD of the feature space. If there were a bag that contained only one positive instance near $P4$ but was still close enough to the negative instances, chances are that this bag will be misclassified as negative. DD defined in such a formulation provides a density-like function that is fickle and is easily affected by introducing even just one negative sample closer to the positive prototype.

The maximization procedure for DD is started from initial guesses. An idea was put forward by Chen and Wang that the maximization should start from every instance in every positive bag (or at least a large sample of positive bags) so that unique local maxima in DD can be identified [9]. A plethora of methods still use this DD formulation [9] [8] [24] [35] [18]. The decision boundary of a DD system is a hyper-ellipsoid in the feature space. A kernel based maximum-margin approach would construct hyperplanar decision boundaries characterizing complex decision surfaces. The first formulation of a support vector machine (SVM) for MIL was proposed in 2002 [2]. They devised an instance-level classifier *mi*-SVM and a bag-level classifier *MI*-SVM. In a way, *MI*-SVM maximized the margin between the most positive instances and the least negative instances in positive and negative bags respectively. The *MI*-SVM framework is now modified and re-christened as *latent*-SVM which plays a central role in the deformable-part models based object recognition algorithms [12]. MILIS provided a similar SVM-based approach with a feedback loop to select instances that provided a higher training stage confidence [13]. This was an idea adapted from a previously existing related idea, MILES [8].

The first distance-based non-parametric, lazy learning approach to MIL was taken by citation-*k*-NN [29]. Inter-bag distances were found using a *minimal* Hausdorff distance. A *k*-nearest neighbour approach was used along with this distance to classify a new bag or to retrieve closer bags. This did not always work in a MIL setting as *k*-NN uses a majority voting scheme. If a positive bag contains fewer number of inherently positive instances than inherently negative instances, majority of its neighbours are going to be negative and the algorithm was confused by the false-positives it reported. Therefore the concept of *citers* was introduced. If *k*-NN *refereed* its neighbours, then its neighbours are *cited* by citers. Citers are the backward propagated references, in the sense that they refer back the considered instance. Though it was a generalized approach, citation *k*-NN did not work as well when positive instances were clustered and such clusters were separated by negative instances, in which case the citers and references did not always compliment each other.

This problem does not apply to all nearest neighbour based approaches. Nearest neighbour approaches should be used properly and their smart usage was discussed in [6]. A novel concept of bag to class (B2C) distance learning was adopted for the use of *k*-NN. A complimentary idea was utilized in a MIL set-up by learning class to bag (C2B) distances by combining all training bags of a particular class to form a super-bag [26] [31]. A similar instance specific distance learning approach was used in [27]. On further study, this was reformulated as a $l_{2,1}$ *minimax* problem and was solved with some effort [28]. A similar idea was implemented to group faces in an image by considering inter bag or bag to bag (B2B) distances in [15]. A related bag to bag approach is used to quantify super-bags in [3].

Most of the MIL algorithms presented above assume that the bags are independent. Though it is a reasonable assumption in a computer vision context, it might not be a general idea. Zhang et al., explored the MIL idea for structured data [34]. A *data-dependent* mixture model approach was developed in [30]. Another approach designed specifically for special data space is the fast bundle algorithm for MIL [5]. One important assumption in the early understanding of MIL is that every positive bag must contain at least one positive instance. Chen et al. felt this was too restrictive and developed a feature mapping using instance selection that projects a MIL problem into a much simpler supervised learning problem using an instance similarity measure [8]. This counter-assumption was also used in a histopathology cancer image learning system using a multiple clustered

instance learning approach [33]. Although in a MIL formulation bag level classification is sufficient and instance level classification though clever, is not required, many algorithms attempt to identify positive instances. A SVM was used to minimize the *hinge* loss (modeled as slack variables) to identify positive instances in [32]. The above methods cater to certain particular configurations of the MIL space and are suitable for particular domains.

3. The proposed approach

Consider figure 2. Though not universal, this figure illustrates a typical MIL feature space. The instances arising from regions $P1$ to $P4$ are potentially inherently positive instances as they are farther away from negative instances while being closer to other positive instances. The instances from positive bags in other regions, along with negative instances are in reality, negative instances as they rub shoulders with negative instances from negative bags.

Suppose we have labeled data $\mathcal{D} = \begin{bmatrix} X^{(1)} & Y^{(1)} \\ X^{(2)} & Y^{(2)} \\ X^{(3)} & Y^{(3)} \\ \vdots & \vdots \\ X^{(n)} & Y^{(n)} \end{bmatrix}$

where $X^{(i)}$ is the i^{th} bag in the dataset and $Y^{(i)} \in \{0, 1\}$ is its label. Internally, each bag $X^{(i)}$ contains m_i (often is a constant m by design, particularly in image classification contexts) instances such that $X^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots x_{m_i}^{(i)}\}$. Consider a small region R of volume V in this feature space. The estimate for the density of instances from positive bags is given by $\frac{(|k^+|)/n}{V}$, where k^+ is the set of instances from positive bags in the region R and $|k^+|$ its cardinality, and n is the number of instances in all of the feature space. Similarly the estimate for the density of negative instances is given by $\frac{(|k^-|)/n}{V}$, where k^- is the set of instances from negative bags in the region R, $|k^-|$ is the number of negative instances in the region R.

Putting them together, $\frac{(|k^+|)/n}{V} - \frac{(|k^-|)/n}{V}$ is a measure that, will be high if the number of positives exceed the number of negatives in that region, will be low if the number of negatives exceed the number of positives in that region, and will be 0 if the number of positives equal the number of negatives within that region. Alternatively, if one considers a (rectangular) Parzen window,

$$\phi(u) = \begin{cases} 1, & |u_j| \leq h \text{ where, } j = 1, 2, ...d, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

the aforementioned measure can also be formulated as,

$$f_{parzen}(x) = \frac{1}{n} \sum_{i=1}^{|k_n^+|} \frac{1}{V} \phi(\frac{x - k_i^+}{h}) - \frac{1}{n} \sum_{i=1}^{|k_n^-|} \frac{1}{V} \phi(\frac{x - k_i^-}{h}) \tag{3}$$

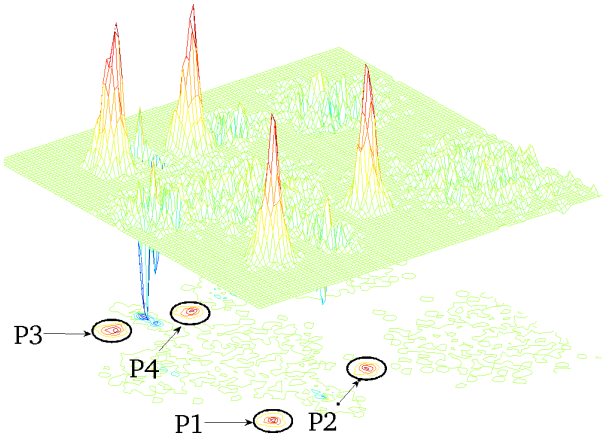

Figure 3. Parsing the MIL feature space with a Parzen window technique. It can be seen that this follows the properties of a MIL density-like.

where, x is any location on the feature space and k_i^+ and k_i^- are instances from positive and negative bags within that region respectively. Such a parsing of the MIL feature space of figure 2 is shown in figure 3. The properties of the function $f_{Parzen}(x)$ hold similar to that of DD and can be easily observed in figure 3. The choice of the size of the region (analogous to the selection of the variance for the Gaussian in the DD formulation) and the Parzen window functions are in line with that of a traditional Parzen window: if the size becomes too large, the measure will not have sufficient resolution. Picking a proper region-size would be a practical difficulty.

Instead of considering a region R of fixed size, let's limit to a fixed number of neighbours k. In this set-up, we start with a region of zero volume from x and grow two regions, one for positive instances and one for negative instances, until we just enclose for each of the regions, k points respectively. This enables us to have different sized regions for positive and negative estimates respectively. While it appears to be a simple k-NN approach to density estimation, we emphasize that we are not using the nearest-neighbour voting rules. In fact, a direct application of nearest-neighbour voting technique will not work on a MIL space as was pointed out by Wang et. al, but the idea of nearest neighbour can still be modified and used to suit the MIL needs [29]. The vote contributions of positive and negative neighbours enclosed by the two regions are their respective kernelized distances to the point x, instead of a uniform majority vote. This aggregated vote can be formu-

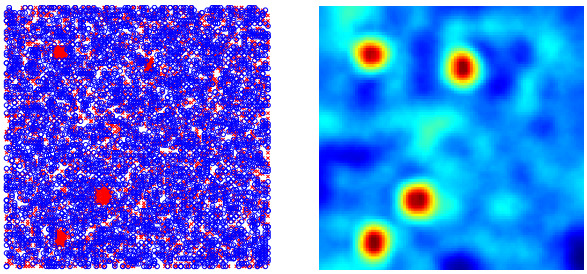

Figure 4. A region of a typical 2D MIL feature space and its parse using the k-NN measure. Red represents positive and blue represents negative.

lated as,

$$f_{kNN}(x) = \sum_{i=1}^{|k^-|} \Psi(||x - k_i^-||) - \sum_{i=1}^{|k^+|} \Psi(||x - k_i^+||)$$
$$\text{such that, } |k^+| = |k^-| = k. \quad (4)$$

where, $\Psi(.)$ is a monotonically increasing sub-modular function, k is the number of neighbours considered, and k^+ and k^- are now the set of k instances from positive and negative bags that are the nearest to x respectively. $\Psi(.)$ is used as a way to scale distances when the featurespace is arbitrarily large. It can be considered as normalization. For all our experiments, we typically use $\Psi(x) = x$.

The advantage of fixing the number of neighbours is that in a region where there are no points or very few number of points, we will get a block of uniform measure and in a region where there is a high density of points, we will get a smoothly varying measure. Such a measure is shown in figure 4. The impact of the number of neighbours k is similar to that of the size of the region R in the Parzen window idea. If k is too small, the measure is going to give information about a very small local region and is thereby unreliable. If k is too large , the impact of proximity is going to be averaged out.

Learning

Learning under this formulation is a straight forward threshold learning and this is done by maximizing the validation accuracy. An instance-level classifier using this measure can be constructed as,

$$h(x) = \mathbb{1}\{f_{kNN}(x) \geq T\} \quad (5)$$

This is an indicator function that outputs 1 if the measure is above a threshold T and 0 if the measure is below the threshold T. We can use this instance-level classifier to construct a bag-level classifier.

$$b(X) = \mathbb{1}\{\sum_{i=1}^{m} h(x_i) \geq a\} \forall x_1, x_2, \ldots, x_m \in X. \quad (6)$$

This is an indicator function that classifies the bag 1 if it has at least a instances classified as positive and 0 other-wise. Typically in most MIL settings $a = 1$, although this need not be the case generally. The aim of this non-parametric empirical risk minimization formulation is to minimize the training error,

$$\epsilon(\hat{b}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{b(X^{(i)}) \neq Y^{(i)}\}, \quad \forall (X, Y) \in D. \quad (7)$$

by estimating $\hat{T}$ that best minimizes $\epsilon(\hat{b})$ as,

$$\hat{T} = \arg\min_{T} \epsilon(\hat{b}) \quad (8)$$

Once the threshold is learnt, classification is performed directly by using the bag-level classifier in equation 6 with the learnt threshold. Note that in MIL, it is not required, although possible in this case, to label each instance in the bag. The labeling of instances can be as follows:

$$y(x) = h(x)|_{T=\hat{T}} \quad (9)$$

This process equivalent to maximizing the equation 15 (or 3) for all points of feature space and considering the local maximas as instance prototypes, as was described by Chen et. al, for the DD formulation [9]. This now enables comparison to prototyping-based methods. Such a formulation can now be re-written as,

$$\hat{x} = \arg\max_{x} \left[\sum_{i=1}^{|k^-|} \Psi(||x - k_i^-||) - \sum_{i=1}^{|k^+|} \Psi(||x - k_i^+||) \right],$$
$$\text{such that,} |k^+| = |k^-| = k. \quad (10)$$

where $\hat{x}$ is a prototype positive instance. One advantage of using equation 10 is that once the prototypes are found, we neither need the entire dataset anymore nor do we need to calculate distances to all the points in the dataset. The prototypes easily divide the featurespace into probabilistic Voronoi tessellations such as in figure 4, or we could estimate a radius around every prototype to isolate hyperspherical regions that are positive.

We solve this optimization problem by using an idea similar to the one used in [9]. We start a local gradient ascent from every instance from every positive bag in the training dataset and find a local maxima. Since such maximas can only ever end in a high density region of true positive instances from positive bags and since we start each gradient ascent from every instance in every positive bag, each ascent is computationally tractable in small number of iterations. Indeed, often few well-chosen instances from positive bags make this convergence faster and such techniques can be found for maximizing the DD in various papers previously surveyed in section 2. Similar techniques can be

applied here as well. All the local maximas are sorted (after non-maximal suppression) and top N are considered as instance prototypes. It is to be noted that for the dataset shown in figure 2, while the top 5 maximas were enough to find all four prototypes for our approach, it takes top 24 maximas for DD to find the four prototypes.

The k for k-NN is picked here by a typical elbow method. Once local maximas (instance prototypes) are found we can again maximize a validation accuracy jointly for all instance prototypes to find a threshold of classification for each prototype in terms of the distance to the prototype, hence creating a hyper-spherical decision regions around each prototype. Thus the decision boundaries of this method creates a tessellation of the feature space. The tessellation is a set of hyper-spherical regions around a positive prototype with varying radii.

4. Experiments and Results

In this section we provide details of our experiments and the results from those experiments. We evaluated our method using three standard MIL datasets: the musk dataset, Andrew's datasets, the Corel datasets (both 1k and 2k), and our own dataset: the DR dataset. For all the results shown on all the datasets, we used the most common implementation methodologies, including data splits, cross validations and average over runs that were found in literature. This enabled us to compare against results that were published in the same. When results were not available or when the protocol doesn't match, we evaluated the results using the codes from CMU MIL toolbox[1]. In case of MILES, the results were obtained by using the author's original code[2]. As per our best knowledge the results provided were obtained for best parameter settings using grid search.

4.1. Musk dataset

An accepted evaluation dataset in the MIL literature is the musk dataset. The musk dataset is well-described in [10]. Musk dataset is a benchmark feature space used to predict drug activity. MUSK 1 contains 92 molecules with 47 musk and 45 non-musk molecules. MUSK 2 contains 102 molecules with 39 musk and 63 non-musk molecules. Each bag contains variable number of instances with 166 dimensional features and binary labels. We use the standard implementation specifications that is used in the original APR paper and other published literature: ten-fold cross-validation over the entire dataset, since its easier to compare against a plethora of methods [10]. Table 1 compares the performance of various algorithms against the proposed method. It can be seen that the proposed method is best in MUSK 1 and among the high performing methods in MUSK 2. MUSK datasets are uni-concept datasets. For instance, in MUSK 1, among a total of 476 unique instances each with feature values ranging from -348 degrees to 336 degrees, there are only 633 unique feature values. In such a heavily quantized feature space that is 166 dimensional, detecting one potential instance prototype is easier for density based algorithms. Our method while being the best in the MUSK 1 data, is also among the better in the MUSK 2 data.

Methods	MUSK 1	MUSK 2
DD [19]	88.9%	82.5%
EM-DD [35]	84.8%	84.9%
citation (k)-NN [29]	**92.4%**	86.3%
mi-SVM [2]	87.4%	83.6%
MI-SVM [2]	77.9%	84.3%
DD-SVM [9]	85.8%	**91.3%**
MILES [8]	86.3%	87.7%
MIforest [31]	85%	82%
MILIS [13]	88.6%	91.1%
ISD [27]	85.3%	79.0%
ALP-SVM [3]	87.9%	86.6%
MIC-Bundle [5]	84%	85.2%
Ensemble [18]	89.22%	85.04%
Proposed	**92.4%**	86.4%

Table 1. Performance of various MIL algorithms on the musk dataset.

4.2. Andrew's datasets

Andrews et. al, in their *mi*-SVM paper proposed the use of three classification datasets, *elephant*, *fox* and *tiger*, for the use of evaluating multiple-instance learning [2]. These are now popular benchmark datasets in the MIL literature. We also test our algorithm on these datasets using the same specifications mentioned on the said article. Each dataset has 200 images with 100 positive and 100 negative images. The number of instances in each category are 1391, 1320 and 1220 respectively with varying number of instances per bag. Each instance is a 230 dimensional feature vector. We train on a $2/3$ random split of the data and test on the remaining $1/3$ of the unseen data. The results are maximized over 15 runs of validation and are shown in table 2. Our result while being the best in the Elephant and Fox classes is almost as good as the best in the Tiger class. It is to be noted that we are significantly higher in the Fox class which is widely considered to be a notoriously noisy dataset for MIL ergo a strong indicator of our method's adaptability.

4.3. Corel dataset

Corel is another well known, image categorization dataset for MIL benchmarking. The Corel-2k dataset consists of 2000 images. There are 20 classes and each class

[1]CMU MIL toolbox: http://www.cs.cmu.edu/~juny/MILL

[2]MILES homepage: http://www.cs.olemiss.edu/~ychen/MILES.html

Methods	Elephant	Fox	Tiger
citation k-NN [29]	79.2%	62.5%	82.6%
mi-SVM [2]	79.7%	62.9%	79%
MILES [8]	70%	56%	62%
MIforest [31]	84%	64%	82%
ISD [27]	77.9%	63%	85.3%
ALP-SVM [3]	84%	69%	**86%**
MIC-Bundle [5]	80.5%	58.3%	79.11%
Ensemble [18]	84.25%	63.05%	79.30%
Proposed	**86%**	**73.94%**	85.7%

Table 2. Performance of various MIL algorithms on Andrew's dataset.

Methods	Corel-1k	Corel-2k
mi-SVM [2]	76.4%	53.7%
MI-SVM [2]	75.1%	55.1%
MILES [8]	82.3%	68.7%
DD-SVM [9]	81.5%	67.5%
MILIS [13]	83.8%	70.1%
Proposed	**87.3%**	**71.9%**

Table 3. Performance of various MIL algorithms on Corel dataset.

consists of 100 images. The Corel-1k dataset is a subset of this dataset with the first 10 difficult categories. Table 3 shows the performance of the proposed approach in the corel dataset. It is to be noted that we are producing the best results in the Corel dataset. Training-testing data is again a $2/3 - 1/3$ split.

4.4. A DR dataset

As was briefly discussed in section 1, DR image classification is an application especially suitable for MIL. In practice, the difficulty in this problem arises from the fact that the physical and observable difference between a normal eye and a pathological eye can be very small, localizing to regions with slightly different characteristics. This can be seen in figure 1.

A variety of classification and retrieval schemes have been tried on DR images. Structural Analysis of the Retina (STARE) is one of the earliest attempts to solve the DR conundrum [21] [14]. STARE performs automated diagnosis and comparison of images to search for images similar in content. Recently other learning approaches were developed to identify relevant patterns using local relevance scores [23]. Application of MIL approaches to DR is gaining interest in recent years [22].

In this study, we consider the auto color correlogram (AuoCC) as a color feature, which is well-studied in the medical imaging literature [16]. A modified and quantized 64-bin AutoCC feature is extracted for each instance in an image. We neglect the black regions and sample 48 non-overlapping instances from every image. We use a high-quality color fundus image database of 425 images comprising 160 normal images, and 265 affected images to test our algorithm on. This dataset was constructed from publicly available databases including DiabRetDB0 [11], DiabRetDB1 [17], STARE [21] and Messidor[3] and has been used in some existing studies [7] [25]. The balance of the database is more towards the positive bags and this makes it more challenging for a MIL algorithm. The results were all evaluated using a $2/3 - 1/3$ train-test split.

Methods	Accuracy
DD [19]	61.29%
EM-DD [35]	73.5%
citation k-NN [29]	78.7%
mi-SVM [2]	70.32%
MILES [8]	71%
Proposed	**81.3%**

Table 4. Performance of various MIL algorithms on DR dataset.

Prototyping DR instances

In the prototyping sense, each prototype of positive instances should roughly correspond to one type of lesion. As we use color features this is easily possible. We estimated a total of about 35 different types of lesion prototypes using our algorithm and verified it with EM-DD's prototypes. EM-DD had its maximum accuracy at about 40 prototypes. It is reasonable to assume from this information that there is somewhere between 35-40 different positive prototypes, each of which in the feature space might correspond to a unique lesion type or character. In this feature space, the negative instances are of three types: normal skin, nerves and the optical disk. This is a reasonably noisy datasets and often has only one or two instances among 48 instances that are positive in a positive bag. Though the distribution of the optic disk might be noisy, and the number of true positive instances are very low, the proposed algorithm has the potential to adjust to it. Table 4 shows the results of the proposed approach on the DR dataset, where the proposed method stands best.

4.5. Sensitivity to labeling error

Although not an implicit feature of the proposal, we perform the experiments to demonstrate the proposed method's sensitivity to labeling error, exactly similar to the one described in [8]. We deliberately flip the labels for a range of percentages of labels randomly on our training split and test the trained model on the original labels in the testing split. The split was $2/3 - 1/3$. The accuracies of the proposed

[3] Kindly provided to us by the messidor program partners. Visit http://messidor.crihan.fr

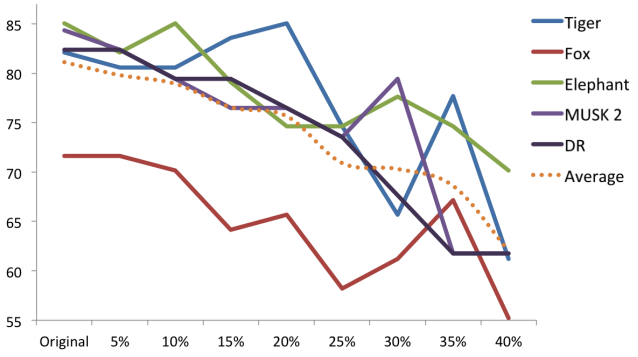

Figure 5. Accuracy vs Percentage of labels flipped for the proposed method. Flatter curve is good.

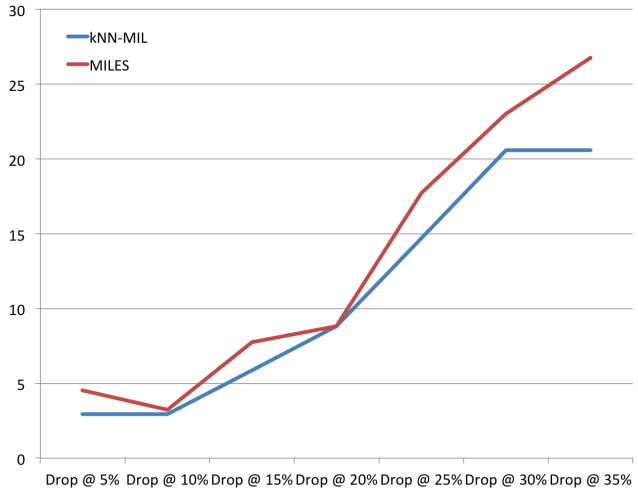

Figure 6. Drop in accuracy at various noise levels for proposed and MILES on the DR dataset. The lower the value the better.

method on various datasets are shown in figure 5. After about 20% of labels are corrupted, the proposed method still loses only about 5% accuracy and only when about one-third of the labels are corrupted, the proposed method loses about 10% accuracy. The average drop in accuracy for both the proposed method and MILES are compared in figure 6. It is clear that MILES and the proposed algorithm follow the exact same trend. This trend is clearly indicative that the proposed method is *as good as* MILES and is often times better, when it comes to sensitivity to labeling noise. It is noteworthy that MILES is considered the state-of-the-art benchmark for sensitivity to labeling error out of all MIL methods published and that was one of its core contributions.

5. Conclusion

In this paper, we postulate whether lazy learning ideas can be carried over from traditional non-parametric methods for supervised learning to a MIL setup. We proposed a simple, yet novel usage of non-parametric learning philosophy to the MIL problem. In particular, we analyzed the MIL feature space using a *k*- NN philosophy and proposed a new formulation based on distances to k-nearest neighbours. The new formulation was compared and contrasted with the widely used DD formulation. The proposed approach was tested on the musk datasets, Andrews dataset and the corel datasets, and was found to be effective. The algorithm was used to solve the DR image classification problem and was found to be the best among other algorithms. We therefore conclude that a non-parametric learning philosophy to MIL not only makes intuitive sense but can also be quite a powerful tool for most general cases.

References

[1] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013. 2

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15:561–568, 2002. 1, 3, 6, 7

[3] B. Antić and B. Ommer. Robust multiple-instance learning with superbags. In *Computer Vision–ACCV 2012*, pages 242–255. Springer, 2013. 2, 3, 6, 7

[4] B. Babenko, M. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE PAMI*, 33(8):1619–1632, 2011. 1

[5] C. Bergeron, G. Moore, J. Zaretzki, C. M. Breneman, and K. P. Bennett. Fast bundle algorithm for multiple-instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1068–1079, 2012. 2, 3, 6, 7

[6] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 3

[7] P. S. Chandakkar, R. Venkatesan, and B. Li. Retrieving clinically relevant diabetic retinopathy images using a multi-class multiple-instance framework. In *SPIE Medical Imaging*, pages 86700Q–86700Q. International Society for Optics and Photonics, 2013. 7

[8] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):1931–1947, 2006. 1, 3, 6, 7

[9] Y. Chen and J. Wang. Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research*, 5:913–939, 2004. 1, 2, 3, 5, 6, 7

[10] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997. 1, 2, 6

[11] T. et al. Diabretdb0: Evaluation database and methodology for diabetic retinopathy algorithms. In *Technical Report*. 7

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 1, 3

[13] Z. Fu, A. Robles-Kelly, and J. Zhou. Milis: Multiple instance learning with instance selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):958–977, 2011. 3, 6, 7

[14] M. Goldbaum, N. Katz, S. Chaudhuri, and M. Nelson. Image understanding for automated retinal diagnosis. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 756. American Medical Informatics Association, 1989. 7

[15] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. *Computer Vision–ECCV*, pages 634–647, 2010. 3

[16] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 762–768. IEEE, 1997. 7

[17] T. Kauppi, V. Kalesnykiene, J. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä. Diaretdb1 diabetic retinopathy database and evaluation protocol. *Proc. Medical Image Understanding and Analysis (MIUA)*, pages 61–65, 2007. 7

[18] Y. Li, D. M. Tax, R. P. Duin, and M. Loog. Multiple-instance learning as a classifier combining problem. *Pattern Recognition*, 46(3):865–874, 2013. 3, 6, 7

[19] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *NIPS*, pages 570–576, 1998. 2, 6, 7, 11

[20] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. In *IEEE ICML*, volume 15, pages 341–349, 1998. 1

[21] B. McCormick and M. Goldbaum. Stare= structured analysis of the retina: Image processing of tv fundus image. In *del USA-Japan Workshop on Image Processing, Jet Propulsion Laboratory, Pasadena, CA*, 1975. 7

[22] G. Quellec, M. Lamard, M. Abràmoff, E. Decencière, B. Lay, A. Erginay, B. Cochener, and G. Cazuguel. A multiple-instance learning framework for diabetic retinopathy screening. *Medical Image Analysis*, 2012. 7

[23] G. Quellec, M. Lamard, B. Cochener, C. Roux, G. Cazuguel, E. Decenciere, B. Lay, and P. Massin. A general framework for detecting diabetic retinopathy lesions in eye fundus images. In *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on*, pages 1–6. IEEE, 2012. 7

[24] R. Rahmani, S. Goldman, H. Zhang, S. Cholleti, and J. Fritts. Localized content-based image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1902, 2008. 2, 3

[25] R. Venkatesan, P. Chandakkar, B. Li, and H. K. Li. Classification of diabetic retinopathy images using multi-class multiple-instance learning based on color correlogram features. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 1462–1465. IEEE, 2012. 7

[26] H. Wang, H. Huang, F. Kamangar, F. Nie, and C. Ding. Maximum margin multi-instance learning. NIPS, 2011. 1, 3

[27] H. Wang, F. Nie, and H. Huang. Learning instance specific distance for multi-instance classification. In *AAAI*, 2011. 2, 3, 6, 7

[28] H. Wang, F. Nie, and H. Huang. Robust and discriminative distance for multi-instance learning. In *IEEE CVPR*, pages 2919–2924. IEEE, 2012. 1, 3

[29] J. Wang and J. Zucker. Solving the multiple-instance problem: A lazy learning approach. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1119–1126. Morgan Kaufmann Publishers Inc., 2000. 1, 3, 4, 6, 7

[30] Q. Wang, L. Si, and D. Zhang. A discriminative data-dependent mixture-model approach for multiple instance learning in image classification,. In *In Proceedings of the 12th European Conference on Computer Vision (ECCV-12),*, 2012. 3

[31] Z. Wang, S. Gao, and L.-T. Chia. Learning class-to-image distance via large margin and l1-norm regularization. In *Computer Vision ECCV 2012*, pages 230–244. 2012. 2, 3, 6, 7

[32] D. Wu, J. Bi, and K. Boyer. A min-max framework of cascaded classifier with multiple instance learning for computer aided diagnosis. In *IEEE CVPR*, pages 1359–1366. IEEE, 2009. 4

[33] Y. Xu, J. Zhu, E. Chang, and Z. Tu. Multiple clustered instance learning for histopathology cancer image segmentation, classification and clustering. CVPR, IEEE, 2012. 4

[34] D. Zhang, Y. Liu, L. Si, J. Zhang, and R. Lawrence. Multiple instance learning on structred data. In *Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS)*, 2011. 3

[35] Q. Zhang and S. Goldman. Em-dd: An improved multiple-instance learning technique. *Advances in neural information processing systems*, 14:1073–1080, 2001. 2, 3, 6, 7

[36] Q. Zhang, S. Goldman, W. Yu, and J. Fritts. Content-based image retrieval using multiple-instance learning. In *Machine Learning-International Worskshop-Then Conference-*, pages 682–689, 2002. 1

6. A simple case-study describing the effectiveness of the proposed method.

In this section we demonstrate by a case-study the strictness of a DD positive neighbourhood.

Consider two bags B^+ and B^- being positive and negative labeled respectively. Consider the instances in the bags as such: $B^+ = \{p, \alpha\}$ and $B^- = \{\alpha, \alpha\}$; such that α be any instance that is so far away form p so that $||p - \alpha|| = \Phi$ where Φ is a large constant and $e^{(-||p-\alpha||)} = 0$. Any instance prototype for a positive instance is therefore at p.

Diverse density at any point x is defined by $f_{DD}(x)$,

$$f_{DD}(x) = \prod_{i=1}^{n} Pr(x = t|B_i^+) \prod_{i=1}^{m} Pr(x = t|B_i^-) \quad (11)$$

for n positive and m negative bags in the data space. Assuming independence between instances, and using the

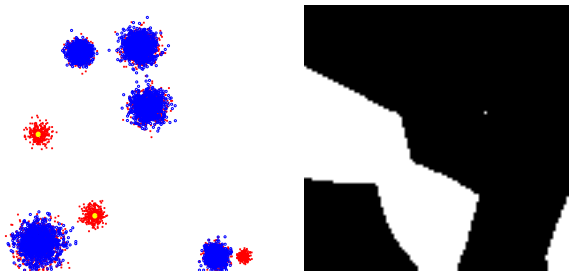

Figure 7. While the EMDD algorithm fails to capture one of the prototypes(left), the proposed method does and classifies that region as positive. Tessellation of the feature space by the proposed method is shown in the right. The accuracy for the proposed method is 100% while the accuracy of EMDD is 77.4%.

noisy-or model, equation 11 can be decomposed to:

$$f_{DD}(x) = [1 - (1 - Pr(x = t|p))(1 - Pr(x = t|\alpha))] \\ (1 - (Pr(x = t|\alpha))^2 \quad (12)$$

DD models the probability $Pr(x = t|i)$ where i be any instance as, $e^{(-||i-x||^2)}$. On the original data space, equation 12 becomes,

$$f_{DD}(x) = [1 - (1 - e^{(-||p-x||)})(1 - e^{(-||\alpha-x||)})] \\ (1 - e^{(-||\alpha-x||)})^2 \quad (13)$$

This equation can be solved at $x = p$ and at $x = \alpha$. At $x = p$ equation 13 becomes, $f_{DD}(p) = [1-(1-1)(1-0)](1-0)^2 = 1$. For the case of N positive and M negative bags with each positive bag containing only one positive instance each, the above measure will be, $f_{DD}(p) = M$ which is also the same case if there were M instances in the one negative bag. Note that DD is not a true density measure. Similarly at $x = \alpha$, $f_{DD}(\alpha) = [1-(1-0)(1-1)](1-1)^2 = 0$, which is also true for the many bags situation. Therefore unless $M = 0$ (no negative bags at all) DD will still be maxima at positive instance prototype.

Adding one additional negative bag with only one instance ($B^* = \{\beta\}$) to the existing database, equation 13 becomes,

$$f^*_{DD}(x) = [1 - (1 - e^{(-||p-x||)})(1 - e^{(-||\alpha-x||)})] \\ (1 - e^{(-||\alpha-x||)})^2(1 - e^{(-||\beta-x||^2)}) \quad (14)$$

At $x = p$, this equation yields, $f^*_{DD}(p) = 1 - e^{(-||\beta-p||^2)}$.

This is a function that is exponentially decreasing in the order of the distance between β and p. The closer the β is to p, the exponentially lower the function is going to become and less the difference will be between, $f^*_{DD}(p)$ and $f^*_{DD}(\alpha)$ as at $x = \alpha$, the equation still remains at $f^*_{DD}(\alpha) = 0$. We also find that $\lim_{\beta\to p} f^*_{DD}(p) = 0$ thus nullifying the prototype as $f^*_{DD}(\alpha)$ is also 0. Although in the strict definition of MIL, such a point is not to be considered a MIL prototype, the belligerent instance could have been noisy. This is truly the case in figure 7 for instance, where due to the presence of a large cluster of negative points next to a positive cluster which contains just one negative point, EMDD cannot identify the positive cluster at all[4].

The proposed MIL formulation on the other hand, learns by threshold learning a function that can also be applied for similar analysis purpose at any point x and can be defined by $f_{kNN}(x)$ as

$$f_{kNN}(x) = \sum_{i=1}^{|k^-|} \Psi(||x - k_i^-||) - \sum_{i=1}^{|k^+|} \Psi(||x - k_i^+||) \\ \text{such that, } |k^+| = |k^-| = k. \quad (15)$$

For the two bag case with $k = 1$, and for $\Psi(a) = a$ the equation becomes, $f_{kNN}(x) = ||x - \alpha|| - ||x - p||$. This equation can be solved at $x = p$ and at $x = \alpha$. At $x = p$, we get $f_{kNN}(p) = ||p - \alpha|| - ||p - p||, = \Phi - 0 = \Phi$, where Φ is the large distance measured between the positive instance p and the negative instance α. For the case of N positive and M negative bags with each positive bag containing only one positive instance each, the above measure will be, $f_{kNN}(p) = (2M - N)\Phi$ which is also the same case if there were M instances in the one negative bag. Notice how unlike DD, where for the case of M negative bags and N negative bags, the value of N didn't feature in $f_{DD}(p)$, our formulation is still dependent on N.

Similarly, at $x = \alpha$, $f_{kNN}(\alpha) = ||\alpha - \alpha|| - ||\alpha - p|| = -\Phi$, a large negative value. For the case of N positive and M negative bags with each positive bag containing only 1 positive instance each, the above measure will be, $f_{kNN}(p) = -N\Phi$. Therefore unless $M = 0$ (no negative bags at all) the proposed approach will still be maxima at positive instance prototype.

Introducing the third bag into the dataset, we get $f_{kNN}(x) = ||x-\alpha|| - ||x-p|| + ||x-\beta||$. At $x = p$, this yields, $f_{kNN}(p) = ||p-\alpha|| - ||p-p|| + ||p-\beta|| = \Phi - 0 + ||p-\beta||$. One can notice here, that as $\lim_{\beta\to p} f_{kNN}(p) = \Phi$ which is the same as the previous case without the negative bag. On the other hand, at $x = \alpha$, the function becomes, $f_{kNN}(\alpha) = 0 - \Phi + ||\alpha - \beta||$. As $\lim_{\beta\to p} f_{kNN}(\alpha) = 0$ which is still Φ, a large value, away from the prototype. Thereby the prototype is still not nullified unlike in the case of DD.

[4]We use EMDD to maximize the DD for instance prototypes in this case.

7. Analogical difference between DD and the proposed formulation.

It is easy to wonder if the proposed formulation is the non-parametric analogy to DD and thus be decomposable from one to another. In this section we attempt to show the fundamental analogical differences between the two formulations and thereby elucidate the philosophical differences between the two.

Consider DD,

$$f_{DD}(x) = \prod_{i=1}^{n} Pr(x = t|B_i^+) \prod_{j=1}^{m} Pr(x = t|B_j^-) \quad (16)$$

where n is the number of positive bags in the dataset and m is the number of negative bags in the dataset. Taking $\log$ we get,

$$f_{lDD}(x) = \log \left[\prod_{i=1}^{n} Pr(x = t|B_i^+) \prod_{j=1}^{m} Pr(x = t|B_j^-) \right] \quad (17)$$

$$= \sum_{i=1}^{n} \log[Pr(x = t|B_i^+)] + \sum_{j=1}^{m} \log[Pr(x = t|B_j^-)] \quad (18)$$

by making an independence (iid) assumption for all the instances in each bag (as is done all through the MIL literature and first introduced in DD itself [19])

$$= \sum_{i=1}^{n} \log[\prod_{k=1}^{|B_i^+|} Pr(x = t|b_{i,k}^+)] + \sum_{j=1}^{m} \log[\prod_{l=1}^{|B_i^-|} Pr(x = t|b_{j,l}^-)] \quad (19)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{|B_i^+|} \log[Pr(x = t|b_{i,k}^+)] + \sum_{j=1}^{m} \sum_{l=1}^{|B_i^-|} \log[Pr(x = t|b_{j,l}^-)] \quad (20)$$

where an instance $b_{i,j}^{+/-}$ is the j^{th} instance from the i^{th} bag and the $+/-$ represents the bag being positive or negative and $|B_r^+|$ and $|B_r^-|$ represents the cardinality of the r^{th} positive and negative bags respectively. Making the substitution $Pr(x = t|a) = \exp(\Psi(||x - a||))$ in the above equation for any instance represented here by a, we get,

$$f_{lDD}(x) = \sum_{i=1}^{n} \sum_{k=1}^{|B_i^+|} \Psi(||x - b_{i,k}^+||) + \sum_{j=1}^{m} \sum_{l=1}^{|B_i^-|} \Psi(||x - b_{j,l}^-||) \quad (21)$$

This equation sums up all the values of all the instances in all the bags, both positive and negative and weights them exponentially. In essence this equation is the energy distribution of all instances positive and negative with respect to x. This equation cannot decompose into equation 15, where negative instances are weighted additively and positive instances are weighted subtractively so as to find regions that are closer to positive instances and farther away from negative instances. The essence of the proposed method is to get away from the DD formulation using bags and to get into the instance space and perform a NN-like instance space tessellation and therein lies the analogical difference between the proposed method and DD.

8. Computational Complexity

From the above discussion on decomposing DD into sums, one can observe that computationally, to estimate functional value at each x (abstracting out the procedure of optimization), the DD form takes $O(no + mq)$ where o is the expected value of number of instances present in a positive bag and q is the expected value of number of instances present in a negative bag. For the k-NN method that complexity is only $O(2k)$, where $k << n$ and $k << m$.

Since estimating time taken for training and testing depends on coding methodologies, choice of optimization solvers, the authors couldn't provide timing information. From the various comparisons that the authors have performed on several synthetic datasets and against the EMDD code authored by Yixen Chen himself, the authors tentatively find the proposed method three times faster than EMDD on average. Similar trends are observed in other datasets as well.

Considering the fact that the proposed method always achieves better accuracy with comparatively lesser number of prototypes than EMDD, just the fact that one has to deal with less number of prototypes also significantly increases the speed during both testing and training time. It is also noteworthy that once after training, we find the prototypes and the radii associated with them, we no longer need to calculate a points' distances using equation 4 all the time, we could simply use the threshold. In this case we only require the prototypes.

9. Sensitivity to k

The number of neighbors k doesn't have as strong an influence on generalization unless the number is too less or too high. The accuracy usually plateaus in a large range of k. All discussions that have been made on choosing a good k for traditional kNN also apply to this formulation as well. We use the "elbow method" to fix a k manually, as mentioned in the paper. We performed an experiment on the MUSK2 dataset by varying k over a large range and plotted

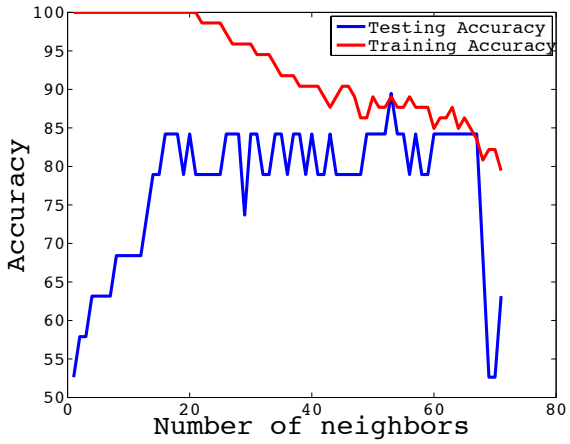

Figure 8. Accuracy vs k. It can be noted that accuracy stabilizes.

the accuracy vs k. The plot is shown in figure 8.

A rule of thumb for picking k is the intuition that you need as many members as half the noisy instances you want to allow around a positive prototype. In this intuition, one may think of choosing k analogous to choosing *slack* in a support vector machine.