2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Model-based Classification and Novelty Detection For Point Pattern Data

Ba-Ngu Vo*, Nhat-Quang Tran*, Dinh Phung† and Ba-Tuong Vo*
*Curtin University, Australia
†Deakin University, Australia

***Abstract*—Point patterns are sets or multi-sets of unordered elements that can be found in numerous data sources. However, in data analysis tasks such as classification and novelty detection, appropriate statistical models for point pattern data have not received much attention. This paper proposes the modelling of point pattern data via random finite sets (RFS). In particular, we propose appropriate likelihood functions, and a maximum likelihood estimator for learning a tractable family of RFS models. In novelty detection, we propose novel ranking functions based on RFS models, which substantially improve performance.**

***Index Terms*—Classification, novelty detection, naive Bayes model, point pattern data, multiple instance data, point process, random finite set.**

I. Introduction

Point patterns are sets or multi-sets of unordered points (or elements) that can be found in numerous data sources. In natural language processing and information retrieval, the 'bag-of-words' representation treats each document as a collection or set of words [1], [2]. In image and scene categorization, the 'bag-of-visual-words' representation–the analogue of the 'bag-of-words'–treats each image as a set of its key patches [3]. In data analysis for the retail industry as well as web management systems, transaction records such as market-basket data [4], [5] and web log data [6] are sets of transaction items. Other examples of point pattern data could be found in drug discovery [7], protein binding site prediction [8].

One simple approach to the classification problem for point patterns is via the naïve Bayes (NB) classifier, see for example [2], [3], [6]. However, the broader task of learning from point pattern data is more appropriately posed as a Multiple Instance Learning (MIL) problem [9], [10], since multiple instance data or 'bags' are indeed point patterns. According to the recent review article [9], there are three paradigms for multiple instance classification, namely Instance-Space (IS), Embedded-Space (EM), and Bag-Space (BS). These paradigms differ in the way they exploit data at the local level (individual data points within each bag) or at the global level (the bags themselves as data points). IS is the only paradigm exploiting data at the local level. At the global level, the ES paradigm maps all point patterns to vectors of fixed dimension, which are then processed by standard classifiers for vectors. On the other hand, the BS paradigm addresses the problem at the most fundamental level by operating directly on the point patterns. The philosophy of the BS paradigm is to preserve the information content of the data, which could otherwise be corrupted through the data transformation process. However, existing methods in the BS paradigm are confined to distance-based approaches [9], while statistical modelling tools and model-based approaches have been overlooked.

In this paper, we introduce statistical models for point pattern data using Random Finite Set (RFS) theory [11], [12], [13]. In particular, we propose appropriate likelihood functions, and a maximum likelihood estimator for learning (from training data) a tractable family of models, called iid-cluster RFSs. Further, in novelty detection where observations are ranked according to their likelihoods, we show that the standard RFS densities are not suitable for point patterns and proposed novel ranking functions that substantially improve performance.

II. A motivating example

The objective of a classifier is to assign a class label $\hat{y} \in \{1, \ldots, C\}$ to an unseen data point X that consists of m feature vectors $x_1, ..., x_m$. In the Bayesian framework, the optimal class label is given by $\hat{y} = \text{argmax}_y p(y \mid X)$, where $p(y \mid X) \propto p(y)\, p(X \mid y)$ is the class posterior probability, $p(y)$ is the prior probability of class y, and

$$p(X \mid y) = p(x_1, ..., x_m \mid y) \tag{1}$$

is the data model or *data likelihood.*

The data model used by the naïve Bayes (NB) classifier [14, pp. 718] imposes a conditional independence assumption among the features so that

$$p(X \mid y) = \prod_{i=1}^{m} p_f(x_i \mid y) \tag{2}$$

where $p_f(x_i \mid y)$ is the conditional probability of the i-th feature given class y (see also [15, pp. 82–89], [16, pp. 380–381]).

In novelty detection or semi-supervised anomaly detection [17], the data likelihood plays an even more important role. In this approach, data are ranked according to a data likelihood (learned from normal data), and data points with likelihoods lower than a threshold are considered as anomalies [18].

Consider the following example on anomalous patterns of daily fallen apples. The apples land on the ground independently from each other, and the probability distribution of the landing positions of the apples is shown in Fig. 1 (the thick solid line). The number of apples and their landing positions

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

for each day are recorded and we are interested in detecting anomalous behavior of the daily apple landing pattern.

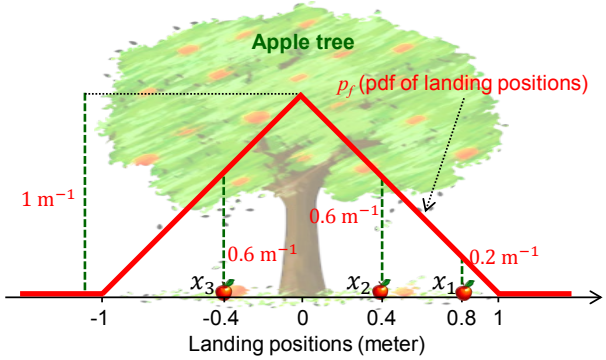


Figure 1. Distribution of landing positions. Position $x_1 = 0.8\,m$ is 3 times less likely than $x_2 = 0.4\,m$ and $x_3 = -0.4\,m$ which are equally likely. Credit: clipartbest.com (apple tree clipart)

Suppose that on day 1 we observed one apple landing at x_1, and on day 2 we observed two apples landing at x_2 and x_3, which of the patterns on these two days is more likely to be an anomaly? To detect anomalies in this setting, it is natural to rank the observed patterns in order of their likelihoods. In this illustration, we follow [2], [3], [6] and use the NB likelihood[1] (2) which gives

$$p(x_1) = p_f(x_1) = 0.2,$$

$$p(x_2, x_3) = p_f(x_2)\, p_f(x_3) = 0.36.$$

where p_f is pdf of landing positions shown in Fig. 1.

Since $p(x_1) < p(x_2, x_3)$, the pattern observed on day 1 *is more likely* to be an anomaly than the pattern observed on day 2. However, if we measure distance in centimeters, then

$$p(x_1) = 0.002 > p(x_2, x_3) = 0.000036,$$

and hence the pattern observed on day 2 *is more likely* to be an anomaly than that of day 1. The likelihood (2) yields contradictory results on the same scenario with different units of measurement!

Note that in the above analysis, the units of the likelihoods were overlooked because $p(x_1)$ is measured in units of m^{-1} or cm^{-1} while $p(x_2, x_3)$ is measured in units of m^{-2} or cm^{-2}. The observed inconsistency arises from the incompatibility in the unit of measurement in the likelihoods $p(x_1)$ and $p(x_2, x_3)$, i.e., we are not "comparing apples with apples". In general, the unit of the likelihood (2) *depends on number of features in* X, i.e. the *cardinality* of X. Hence, comparing the likelihood (2) of different observations is not meaningful, unless they have the same number of features or the features are intrinsically unitless.

Apart from the inconsistency with unit of measurement, such point pattern likelihood also suffers from another problem associated with cardinality. Let us revisit the fallen apples example, however to eliminate the effect of the unit mismatch, this time we restrict ourselves to a finite number of landing positions, by discretizing the interval $[-1\,\mathrm{m}, 1\,\mathrm{m}]$ into 201 points $\{-100, \ldots, 100\}$ and round the landing positions to the nearest of these points (Fig. 2). Thus, instead of a probability density of the landing positions on the interval $[-1\,\mathrm{m}, 1\,\mathrm{m}]$ we now have a (unitless) probability mass function (pmf) on the discrete set $\{-100, \ldots, 100\}$, see Fig. 2.

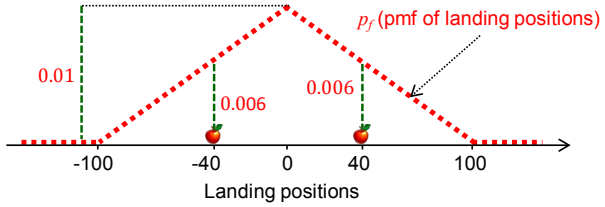
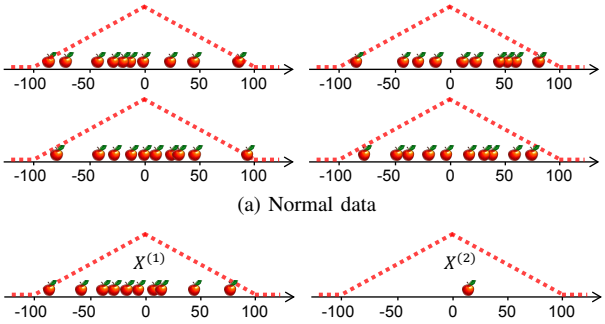


Figure 2. Distribution of discrete landing positions.

Fig. 3a shows 4 'normal' patterns of fallen apples, each with about 10 locations i.i.d. from the pmf of Fig. 2. Two new observations $X^{(1)}$ and $X^{(2)}$ whose features are also i.i.d. from the same pmf are shown in Fig. 3b.



(a) Normal data

(b) New observations. Note that, by NB likelihood, we have $p(X^{(1)}) \approx 2 \times 10^{-23}$ and $p(X^{(2)}) = 0.009$

Figure 3. Examples of normal data and anomaly.

Since $X^{(2)}$ has only 1 feature whereas the 'normal' observations each has around 10 features, it is intuitively obvious that $X^{(2)}$ is anomalous. However, its likelihood is much higher than that of $X^{(1)}$ – a normal datum (0.009 versus 2×10^{-23}). This counter intuitive behavior cannot be attributed to the measurement unit inconsistency because the pmf of the features is unitless.

III. Models for Point Pattern Data

The likelihood (2) was used in the above discussions to illustrate discrepancies with measurement unit and cardinality. However, these discrepancies arise even in the full joint likelihood (1). In this section, we propose models for point pattern data using random finite set, which could address these issues.

A. Random Finite Set

Point patterns can be modeled as random finite sets (RFSs), or simple finite point processes. Point process theory, in general, is concerned with abstract random counting measures. RFSs are geometrically more intuitive and thus better suited for the type of discussions in this article. The likelihood of a point pattern of discrete features is straightforward since

[1]For compactness, the condition on the normal class is omitted, i.e., $p(X)$ is used instead of $p(X \mid y = \text{'normal'})$.

this is simply the product of the cardinality distribution and the joint probability of the features given the cardinality. The difficulties arise in continuous feature spaces. In this work, we only consider continuous feature spaces.

Let $\mathcal{F}(\mathcal{X})$ denote the space of finite subsets of a space $\mathcal{X}$. A random finite set (RFS) X of $\mathcal{X}$ is a random variable taking values in $\mathcal{F}(\mathcal{X})$ [19], [11], [12], [20], [13]. In essence, an RFS is a finite-set-valued random variable that is random in the number of elements, as well as the values of the elements. An RFS X can be completely specified by a discrete (or categorical) distribution that characterizes the cardinality $|X|$, and a family of symmetric joint distributions that characterizes the distribution of the points (or features) of X, conditional on the cardinality.

Analogous to random vectors, the probability density of an RFS (if it exists) is essential in the modeling of point pattern data. The probability density $p : \mathcal{F}(\mathcal{X}) \to [0, \infty)$ of an RFS is the Radon-Nikodym derivative of its probability distribution relative to the dominating measure μ, defined for each (measurable) $\mathcal{T} \subseteq \mathcal{F}(\mathcal{X})$, by [19], [12], [21], [22]:

$$\mu(\mathcal{T}) = \sum_{m=0}^{\infty} \frac{1}{m! U^m} \int \mathbf{1}_{\mathcal{T}}(\{x_1, ..., x_m\})\, d(x_1, ..., x_m) \quad (3)$$

where U is the unit of hyper-volume in $\mathcal{X}$, and $\mathbf{1}_{\mathcal{T}}(\cdot)$ is the indicator function for $\mathcal{T}$. The measure μ is the unnormalized distribution of a Poisson point process with unit intensity $u = 1/U$ when $\mathcal{X}$ is bounded. Note that μ is unitless and consequently the probability density p is also unitless.

In general the probability density of an RFS, with respect to μ, evaluated at $X = \{x_1, ..., x_m\}$ can be written as

$$p(X) = p_c(m)\, m!\, U^m f_m(x_1, ..., x_m), \quad (4)$$

where $p_c(m) = \Pr(|X| = m)$ is the cardinality distribution, and $f_m(x_1, ..., x_m)$ is a symmetric joint probability density of the points $x_1, ..., x_m$ given the cardinality, see [23, p. 27] ((Eqs. (1.5), (1.6), and (1.7)), [12], [21].

B. Likelihoods for point pattern data

Instead of a random vector X, we propose to model each point pattern as an RFS X. A general form for the likelihood of X is given by (4), which can capture the cardinality information as well as the dependence between the features. The RFS data model also avoids the unit of measurement inconsistency since the probability density with respect to μ is unitless.

Imposing the 'naïve' conditional independence assumption among the features on the model in (4) reduces to the *iid-cluster RFS* model [11]

$$p(X) = p_c(|X|)\, |X|!\, [U p_f]^X \quad (5)$$

where p_f is a probability density on $\mathcal{X}$, referred to as the *feature density,* and $h^X \triangleq \prod_{x \in X} h(x)$, with $h^{\emptyset} = 1$ by convention, is the finite-set exponential notation.

When p_c is a Poisson distribution we have the celebrated *Poisson point process* (aka, *Poisson RFS*)

$$p(X) = \lambda^{|X|}\, e^{-\lambda}\, [U p_f]^X \quad (6)$$

where λ is the mean cardinality. The Poisson model is completely determined by the intensity function $u = \lambda p_f$ [12],[21], [22]. Note that the Poisson cardinality distribution is described by a single non-negative number λ, hence there is only one degree of freedom in the choice of cardinality distribution for the Poisson model.

Given the training data, a key task in learning is to compute estimates of the underlying parameters of the model. Learning the general model (4) is computationally intensive. The iid-cluster model (5), on the other hand, provides a good trade-off between tractability and flexibility.

C. Maximum Likelihood Estimation

This subsection presents a solution to learning the parameters of an iid-cluster RFS model using maximum likelihood (ML) estimation.

Given a finite list of observations $Z^{(1)}, ..., Z^{(N)} \in \mathcal{Z}$ and a parametrized probability density $f(\cdot \mid \theta)$ on $\mathcal{Z}$, we denote

$$\hat{\theta}(f; Z^{(1)}, ..., Z^{(N)}) \triangleq \underset{\theta}{\operatorname{argmax}} \left(\prod_{n=1}^{N} f(Z^{(n)} \mid \theta) \right) \quad (7)$$

If the data points $Z^{(1)}, ..., Z^{(N)}$ are i.i.d. according to $f(\cdot \mid \theta)$, then $\hat{\theta}(f; Z^{(1)}, ..., Z^{(N)})$ is indeed the maximum likelihood estimate (MLE) of θ.

Since an iid-cluster RFS is uniquely determined by its cardinality and feature distributions, we consider cardinality and feature distributions parametrized by $p_c(\cdot|\theta_c)$ and $p_f(\cdot|\theta_f)$, i.e.

$$p(X \mid \theta_c, \theta_f) = p_c(|X| \mid \theta_c)\, |X|!\, U^{|X|}\, [p_f(\cdot|\theta_f)]^X \quad (8)$$

Learning the underlying parameters of an iid-cluster model amounts to estimating $\theta = (\theta_c, \theta_f)$ from training data. Furthermore, the MLE of the iid-cluster model parameters separates into the MLE of the cardinality distribution parameters θ_c and MLE of the feature density parameters θ_f. This is stated more concisely in the following Proposition.

Proposition 1. Let $X^{(1)}, ..., X^{(N)}$ be N i.i.d. realizations of an iid-cluster RFS with parametrized cardinality distribution $p_c(\cdot|\theta_c)$ and feature density $p_f(\cdot|\theta_f)$. Then the MLE of (θ_c, θ_f), is given by

$$\hat{\theta}_c = \hat{\theta}\left(p_c; |X^{(1)}|, ..., |X^{(N)}|\right) \quad (9)$$

$$\hat{\theta}_f = \hat{\theta}\left(p_f; \uplus_{n=1}^{N} X^{(n)}\right) \quad (10)$$

where $\uplus_{n=1}^{N} X^{(n)}$ is the disjoint union of $X^{(1)}, ..., X^{(N)}$.

Proof. Using (8), we have

$$\prod_{n=1}^{N} p(X^{(n)} \mid \theta_c, \theta_f)$$
$$= \prod_{n=1}^{N} p_c(|X^{(n)}| \mid \theta_c)\, |X^{(n)}|!\, U^{|X^{(n)}|} \prod_{x \in X^{(n)}} p_f(x \mid \theta_f)$$
$$= \left(\prod_{n=1}^{N} p_c(|X^{(n)}| \mid \theta_c) \right) \cdot \left(\prod_{n=1}^{N} |X^{(n)}|!\, U^{|X^{(n)}|} \right)$$
$$\cdot \left(\prod_{n=1}^{N} \prod_{x \in X^{(n)}} p_f(x \mid \theta_f) \right) \tag{11}$$

Hence, to maximize the likelihood we simply maximize the first and last bracketed terms in (11) separately. This is achieved with (9) and (10). QED.

Observed from Proposition 1 that the MLE of the feature density parameters is identical to that used in NB. For example, if the feature density is a Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the parameters of its ML estimates are:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{x \in X^{(n)}} x, \tag{12}$$
$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{x \in X^{(n)}} (x - \hat{\boldsymbol{\mu}})(x - \hat{\boldsymbol{\mu}})^T. \tag{13}$$

Consequently, the iid-cluster model requires only one additional task of computing the MLE of the cardinality distribution parameters, which is relatively inexpensive.

For a categorical cardinality distribution, i.e. $\theta_c = (p_1, ..., p_K)$, the MLE of the cardinality distribution is given by

$$\hat{p}_k = \frac{1}{N} \sum_{n=1}^{N} \delta_k(|X^{(n)}|). \tag{14}$$

Since there are K parameters $p_1, ..., p_K$, we require a sufficiently large dataset (significantly larger than K). For a small dataset, a cardinality distribution with a small number of parameters should be used to avoid over-fitting, e.g. Poisson, i.e. $\theta_c = (\lambda)$ where its MLE is given by

$$\hat{\lambda} = \frac{1}{N} \sum_{n=1}^{N} |X^{(n)}|. \tag{15}$$

Conceptually, Proposition 1 can be extended to the general RFS model (4), which relaxes the naïve independence assumption. The MLE of the cardinality distribution parameters is computed as for the iid-cluster RFS model. However, for the feature distribution, instead of $\hat{\theta}(p_f; \uplus_{n=1}^{N} X^{(n)})$ we need to compute the MLE of the parameters for the joint densities, i.e.

$$\underset{\theta}{\operatorname{argmax}} \prod_{n=1}^{N} f_{|X^{(n)}|}(X^{(n)} \mid \theta_f) \tag{16}$$

which is far more complex in general. Imposing additional assumptions such as TAN may provide some simplifications. Alternative models such as mixture of iid-cluster RFSs [24] are also promising. However, these are topics for future research.

D. Numerical experiments

This subsection presents two classification experiments with simulated and real-world data. These experiments involve learning from training data, and then use the learned models to classify new observations. The results are bench-marked against the NB model. The first experiment uses simulated data so as to illustrate the benefit of cardinality information when the features between the classes are similarly distributed. The second experiment uses real-world data from the Texture images dataset [25].

1) Classification with simulated data: In this experiment, data are simulated from an underlying model consisting of three clusters C_1, C_2 and C_3. A datum from C_i is a finite set X whose cardinality is Poisson distributed with mean λ_i, and whose features are i.i.d. from a 2-D Gaussian $\mathcal{N}(\cdot; \mu_i, \Sigma_i)$ where

$$\lambda_1 = 6, \quad \mu_1 = [1, 2]^T, \quad \Sigma_1 = \text{diag}[20, 40],$$
$$\lambda_2 = 15, \quad \mu_2 = [2, 3]^T, \quad \Sigma_2 = \text{diag}[60, 20],$$
$$\lambda_3 = 30, \quad \mu_3 = [2, 2]^T, \quad \Sigma_3 = \text{diag}[30, 30].$$

Fig. 4 plots the features and cardinalities of data sampled from these models.

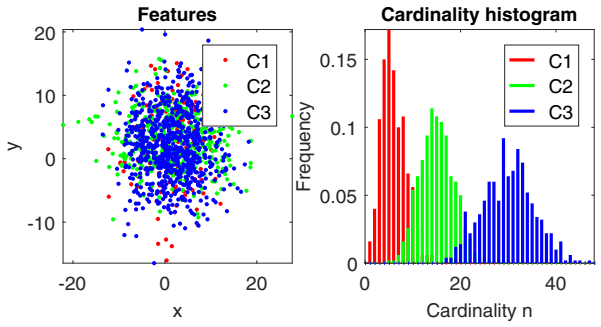
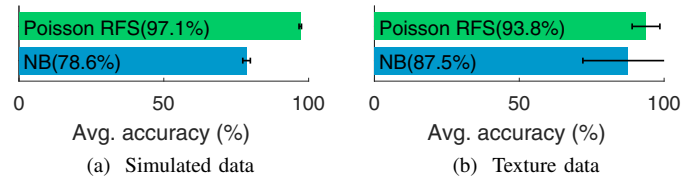


Figure 4. Simulated data for the experiment in section III-D1. Left: Features of the data. Right: Cardinality histogram of the data.

Both the NB and RFS models are trained via ML estimation on a fully observed training dataset set consisting of 900 data points (300 per cluster). For NB, the data models are three 2-D Gaussians (one for each cluster). For the RFS models, we use three Poisson RFSs with 2-D Gaussian feature distributions (which are in fact the same as the Gaussians learned from the NB models).

After training, we evaluate the learned models by classifying a test set consisting of 1500 data points (500 per cluster, also simulated from the described model). The evaluation are run 10 times with 10 different test sets, and the average accuracies[2] are shown in Fig. 5a. Observed that the RFS model can exploit the cardinality information and delivers better performance than NB.



(a) Simulated data (b) Texture data

Figure 5. Performance of classification by NB and Poisson RFS. Note that the error-bars (on the peak of each column) are standard deviations of accuracies, which show that Poisson RFS works more stably than NB.

[2] Accuracy $\triangleq$ $\frac{\text{No. of correct classifications}}{\text{No. of observations in the test set}}$

2) Classification with real data: The second experiment involves classification of the two classes "T14_brick1" and "T15_brick2" from Texture images dataset [25]. Fig. 6 show some example images from these classes.

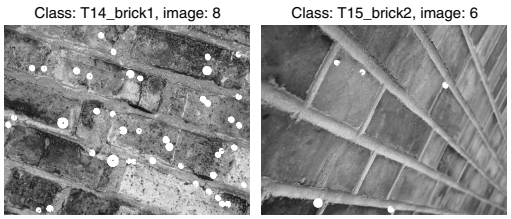


Figure 6. Example images from 2 classes "T15_brick2" and "T14_brick1" in Texture dataset. Circles mark detected SIFT keypoints.

Features are extracted from each image by applying the SIFT algorithm[3], followed by Principal Component Analysis (PCA) to convert the 128-D SIFT features into 2-D features. Thus each image is compressed into a point pattern of 2-D features. Fig. 7 plots the 2-D features of the images in the Texture images dataset.

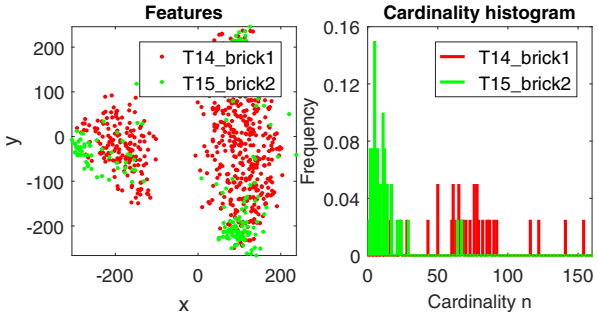


Figure 7. Extracted data from images of the Texture dataset. Left: 2-D features (after applying PCA to the SIFT features). Right: histogram of cardinalities of the extracted data.

The model parameters are learned by using MLE on a training dataset containing 30 images per class. For the NB model, we use 2-D Gaussian distributions to model the data; and for the RFS model, we use Poisson RFSs with 2-D Gaussian feature distributions. After training, the learned models are evaluated on a test set containing all remaining images of the classes (10 images per class). The performance is evaluated with 4-fold cross validation, and the average results are shown in Fig. **??**. Observe that the RFS model outperforms NB, since it can exploit cardinality information of the data.

IV. Ranking of Data

The previous section shows that RFS models avoid the unit inconsistency and improve the classification performance by exploiting cardinality information in point pattern data. However, using the RFS density as the data ranking function for novelty detection does not result in a good performance[4]. Due to the non-uniformity of the reference measure, the probability densities for RFSs presented in the previous section do not provide a consistent ranking of the data. In this section we propose a consistent ranking function for the iid-cluster model.

A. Ranking Function

Note that the probability density of an iid-cluster RFS (5) is a product of p_f^X and a term that depends only on the cardinality $p_c(|X|)\,|X|!U^{|X|}$. Given the cardinality, p_f^X is the density of the feature distribution relative to the Lebesgue measure, and hence is the likelihood of X given its cardinality. However, $p_c(|X|)\,|X|!U^{|X|}$ is proportional to the Radon-Nikodym derivative of the cardinality distribution relative to a Poisson distribution, and thus does not indicate how likely the cardinality of X is. A ranking function ℓ that can reconcile this problem is:

$$\ell(X) \propto p_c(|X|)\,[Cp_f]^X \tag{17}$$

where C is an unknown constant.

To determine a suitable C, consider first the special case where p_f is uniform on a bounded state space $\mathcal{X}$. In such case, all finite set subsets of $\mathcal{X}$ with the same cardinality are equally likely, and a consistent ranking function should satisfy $\ell(X) \propto p_c(m)$, given $|X| = m$. This condition can be generalized to non-uniform p_f by replacing $\ell(X)$ with its expected value, given $|X| = m$

$$\mathbb{E}_X\left[\ell(X) \mid |X| = m\right] \propto p_c(m) \tag{18}$$

Combining (17) with (18) and using the i.i.d. property of the features in iid-cluster models yields:

$$\ell(X) \propto p_c(|X|)\left[\frac{p_f}{||p_f||_2^2}\right]^X \tag{19}$$

where $||p_f||_2^2 = \int p_f^2(x)dx$ is the squared L^2-norm of p_f, which has units of U^{-1}, and hence $\ell(X)$ is unitless.

B. Numerical experiments

In this section we compare the performance of the proposed ranking function (19) with the NB likelihood and Poisson RFS likelihood in novelty detection. The threshold is set at the 2^{nd} 10-quantile of the ranking function values[5] of the training dataset (consisting of only normal data). Observations ranked below this threshold are classified as anomalies.

1) Novelty detection with simulated data: In this experiment, normal data are samples, having cardinalities between 40 and 60, drawn from a Poisson RFS with mean cardinality 48 and a 2-D Gaussian feature distribution

$$\mathcal{N}\left(\cdot; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.06 & 0.01\\0.01 & 0.04\end{bmatrix}\right).$$

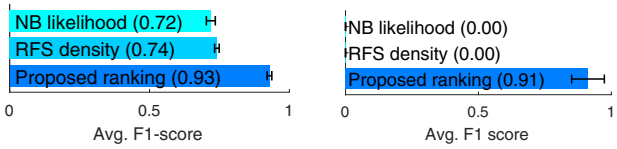
The same dataset containing 500 normal data points is used to train the NB and Poisson RFS models with Gaussian feature distributions via MLE (subsection III-C).

[3]Using the VLFeat library [26].

[4]We have done several experiments which are not shown here due to the paper's length limitation.

[5]The performance depends on the (manually selected) threshold. Nonetheless, the performance of the proposed ranking function would still be better than the others since it can rank the data consistently as shown in the boxplots of likelihood values for the experiment.

Three types of anomalies are considered: *low-cardinality anomaly* (cardinality ≤ 10), *high-cardinality anomaly* (cardinality ≥ 80), and *feature anomaly* which has the same cardinality with normal data, but the Gaussian feature distribution now has a mean of $[1, 1]^T$. Novelty detection is performed on a test set containing 200 normal observations and 300 anomalies (100 for each type). The tests are run 10 times with 10 different randomly sampled test sets. The averaged results shown in Fig. 8a indicated that the proposed ranking function yields superior performance.



(a) Simulated data (b) Texture data

Figure 8. Novelty detection results by 3 ranking functions: NB likelihood, Poisson RFS likelihood, and proposed ranking function.

2) Novelty detection with real data: This experiment uses the real dataset from the second experiment in subsection III-D. Normal data are taken from the "T14_brick1" class and anomalous test data are taken from the "T15_brick2" class. A 4-fold cross-validation is used: for training we use 30 images from the normal class and for testing we use the remaining images from normal class (10 images) and 10 images from abnormal class (different at each time).

As shown in Fig. 8b, ranking the data using the NB likelihood and Poisson RFS likelihood could not detect any anomalies, while the proposed ranking function achieved an F1 score[6] around 0.91. Moreover, Fig. 9 verified that only the proposed ranking function provides a consistent ranking, while the NB and Poisson RFS likelihoods even rank anomalies higher than normal data.

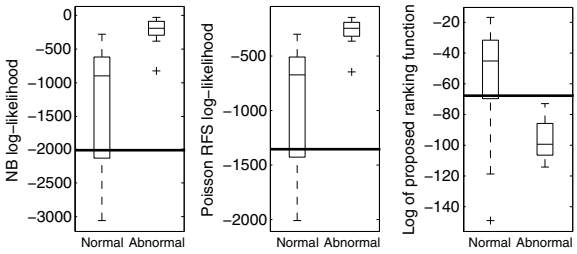


Figure 9. Boxplots of likelihoods computed by 3 models, namely NB likelihood, RFS density, and the proposed ranking function, for a fold of Texture dataset. The solid line going through each graph is the threshold (the 2^{nd} 10-quantile).

V. Conclusions

In this paper, we have introduced statistical models for point pattern data using Random Finite Set theory. Such models provide the means for developing model-based classification and novelty detection for point pattern data. In particular we proposed a maximum likelihood method for learning the parameters of an iid-cluster RFS–the analogue of the naïve Bayes model for point patterns. For novelty detection, we proposed novel ranking functions based on RFS models, which substantially improve performance. Our results also contribute to the Bag-Space paradigm in multiple instance learning where statistical models are not available.

[6] *F1 score* $\triangleq 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

References

[1] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." DTIC Document, Tech. Rep., 1996.

[2] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, 2004.

[4] S. Guha, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," in *Data Engineering, 1999. Proceedings., 15th International Conference on*. IEEE, 1999, pp. 512–521.

[5] Y. Yang, X. Guan, and J. You, "Clope: a fast and effective clustering algorithm for transactional data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 682–687.

[6] I. V. Cadez, S. Gaffney, and P. Smyth, "A general probabilistic framework for clustering individuals and objects," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 140–149.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[8] F. Minhas and A. Ben-Hur, "Multiple instance learning of calmodulin binding sites," *Bioinformatics (Oxford, England)*, vol. 28, no. 18, pp. i416–i422, 2012.

[9] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[10] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, no. 01, pp. 1–25, 2010.

[11] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes*. Springer, 1988, vol. 2.

[12] J. Moller and R. P. Waagepetersen, *Statistical inference and simulation for spatial point processes*. Chapman & Hall CRC, 2003.

[13] R. P. Mahler, *Advances in statistical multisource-multitarget information fusion*. Artech House, Inc., 2014.

[14] S. Russell and P. Norvig, *Artificial intelligence: A modern approach*. Prentice Hall, 2003.

[15] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[16] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[17] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[18] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[19] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 1995.

[20] R. P. Mahler, *Statistical multisource-multitarget information fusion*. Artech House, Inc., 2007.

[21] B.-N. Vo, S. Singh, and A. Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 41, no. 4, pp. 1224–1245, 2005.

[22] H. G. Hoang, B.-N. Vo, B.-T. Vo, and R. Mahler, "The Cauchy-Schwarz divergence for Poisson point processes," *IEEE Trans. Information Theory*, 2015.

[23] M. van Lieshout, *Markov point processes and their applications*. Imperial College Press, 2000.

[24] D. Phung and B.-N. Vo, "A random finite set model for data clustering," in *Proc. 17th Annu. Conf. Information Fusion*, Salamanca, Spain, 2014.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1265–1278, 2005.

[26] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.