IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 14, 2020, accepted December 22, 2020, date of publication December 25, 2020,
date of current version January 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047536

Efficient Feature Mapping in Classifying Proportional Data

MD. HAFIZUR RAHMAN AND NIZAR BOUGUILA, (Senior Member, IEEE)
Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC H3G 1M8, Canada

Corresponding author: Nizar Bouguila (nizar.bouguila@concordia.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

ABSTRACT In image classification, traditional kernels or feature mapping functions of Support Vector Machine(SVM) use discriminative features without considering the true nature of the data. Our work in this paper is motivated by the need to consider intrinsic distribution of *L1* normalized histograms and develop a flexible feature mapping technique by combining histogram based features and distribution based density features. The proposed mapping technique contains prior knowledge about the the data which provides a flexible representation and thus increases the discriminative power of the classifier. Such flexibility is achieved due to the explanatory capabilities of Dirichlet, generalized Dirichlet and Beta-Liouville distributions to model proportional data. In addition to that, we present a general framework to estimate the parameters of these distributions by taking maximum likelihood (MLE) approach. Experimental results show that the proposed technique increases the effectiveness of SVM kernels for different computer vision tasks such as natural scene recognition, satellite image classification and human action recognition in videos.

INDEX TERMS Proportional data, support vector machines, Dirichlet distribution, generalized Dirichlet distribution, Beta-Liouville distribution, human action recognition, image classification.

I. INTRODUCTION

Appropriate and accurate representation of the data for classification models is one of the existing problems in machine learning. Several classification and hybrid models have been developed. However, a little attention has been given to a get a proper representation of the data through distribution based feature mapping in discriminative approaches [1]. In this paper, we address this issue in supervised learning problems for proportional data. A popular image representation is the Bag of Visual Words (BoVW) which is essentially quantizing similar patches of an image to the corresponding cluster center which is known as codebook [2], [3]. Modelling such data after normalization in a probabilistic manner needs to satisfy the constraints of non-negativity and unit sum. Examples of such data includes *L1* normalized histogram for images and normalized bag of words representation of texts (or images) data. In particular, we are motivated by the problem of modelling features in images and videos where each feature represents a portion of the total features considered. For example, an image can be represented by a normalized histogram of bag of vectors where each vector element represents a sub-region of the image. Knowledge about statistical characteristics of such representations has to be used effectively in order to get better accuracy for the classification tasks. Dirichlet, Generalized Dirichlet and Beta-Liouville distributions can model this type of data to get the prior information which can be used as a feature. The advantage of such distributions are that they can capture the nature of the data and provide flexibility. For SVM, traditional kernels do not take into account the nature of the data. Incorporating our proposed feature mapping technique increases the classification accuracy of these kernels.

Performance of machine learning algorithms depend on the representation of the input data. Incorporating invariance in the representation using prior knowledge is a common technique to make the learning task more efficient and in general, prior information makes it possible to generalize training examples to novel test examples [1]. In supervised learning, hyperparameters of the classifiers work as prior information. Another approach is to select features that convey most relevant information regarding the data or the task. Such features are automatically incorporated in some kernels such polynomial kernel for SVM [4]. On a different note, distribution based flexible feature mapping can be efficient in different classification tasks [5]. Our contribution falls into

The associate editor coordinating the review of this manuscript and approving it for publication was Bohui Wang.

the second category. For SVM, input data are represented as points in high dimensional space. This representation needs to be linearly separable to make the model work properly. Therefore, for non-linear data, performance of SVM model lacks accuracy. However, kernel trick or feature mapping technique has made it possible to model non-linear data which is essentially taking the data space to higher dimension where the data become linearly separable. It is a common idea to extract new features from the input variables through a feature mapping function which increases the separability between the data classes. On the contrary, feature mapping without statistical measure about the data does not guarantee the improvement in model's performance. Selecting the most informative attributes from the set of redundant attributes is sub-optimal for a classifier and on the contrary, it may keep out some relevant features as well [6]. Therefore, extracting or creating new features from the data with prior information using a parameterized feature mapping function can be incorporated in classification model with certain degree of confidence. Histogram representation of the extracted data can be modelled in a probabilistic way by performing *L1*-normalization and Dirichlet or Liouville type distributions is the choice to estimate the density of such data. Therefore, a parametric distribution based mapping function can be developed to increase the flexibility of the datapoints in the feature space.

Rest of the paper is organized as follows, section II highlights some related works, section III introduces the Dirichlet, generalized Dirichlet and Beta-Liouville distributions along with the parameter estimation method for these distributions. Support vector machine and kernel tricks are discussed in section IV. Our proposed feature mapping technique is discussed in section V. In section VI, we show the experimental results of the proposed methodology for image and video classification tasks. Concluding remarks are discussed in section VII.

II. RELATED WORKS

Many researchers have focused on improving traditional SVM by implementing new kernels or new feature mapping functions. To teach the machine to differentiate between different images, frequencies of local features of an image or video frame are quantized into a histogram [7]–[9]. Feature mapping function proposed by [5] based on Dirichlet distribution has proved to be efficient in different classification and regression tasks for proportional data. In addition, the authors indicated data normalization technique for proportional data to be used in the feature mapping function. Histogram based feature transformation with probabilistic modelling is addressed in [10]. Kernel based methods have been applied in different learning tasks such as Gaussian kernels with different distance measures which proved to be efficient in image classification task [11]. In addition, a combination of generative and probabilistic learning is shown to be effective in image recognition and segmentation tasks [12], [13]. In such approaches the kernel is generated by learning the generative process of the data using probabilistic models such as Gaussian mixture model [14] or Dirichlet and related distributions based mixture models [12], [15]–[17]. In contrast to this, we consider a feature mapping function which considers both discriminative features and density features. In supervised learning, measuring the similarity of $L1$ normalized histograms using Euclidean distance is not effective [11]. In such cases, histogram distances such as χ^2 distance has proved to be effective [18]–[21]. Several histogram based distances and their derivatives have been proposed by many researchers such as [22], [23], [24]–[26]. Apart from this, [27] proposed a non-linear mapping technique based on polar coordinate system. Modification of RBF kernel using first order Taylor series approximation proposed by [28] has achieved better accuracy for semanteme data. However, in contrast to these approaches, we are interested in increasing the discriminative power of SVM using more flexible feature mapping technique for proportional data.

III. DISTRIBUTIONS FOR PROPORTIONAL DATA

A. DIRICHLET DISTRIBUTION

Dirichlet distribution is the generalization of Beta distribution and most appropriate candidate in probability and statistics when modelling proportional data [29]. It is a distribution over the multinomials in a simplex with supports [0, 1]. If a vector $p = (p_1, p_2, \ldots, p_D)$ of length D resides in a D dimensional closed simplex of $\mathbb{R}^D$ then it is defined as,

$$\mathbb{C}(1) = \{p \in \mathbb{R}^D : p_1 + p_2 + \ldots + p_D = 1; \quad p_d \geq 0, 1 \leq d \leq D\} \tag{1}$$

If the proportional vector $p \in \mathbb{C}(1)$,[1] then the joint probability density function of $p = (p_1, p_2, \ldots, p_D)$ is defined as,

$$P(p|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_d \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} p_d^{\alpha_d - 1}$$

$$\sum_{d=1}^{D} p_d = 1, \quad p_d \geq 0 \tag{2}$$

where Γ denotes the gamma function and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_D)$ is a positive parameter vector which defines the shape of the distribution in D dimensional space. Total mass, $\alpha_0 = \sum_d \alpha_d$ is the concentration or scale parameter and the base measure $(\alpha'_1, \alpha'_2, \ldots, \alpha'_D) = \frac{\alpha_d}{\alpha_0}$. In case of symmetric distribution, the mean of the distribution is determined by the base measure. In addition, altering the measurements in $\boldsymbol{\alpha}$ affects the variance of the distribution.

$$E(p_d) = \frac{\alpha_d}{\alpha_0} = \alpha'_d$$

$$Var(p_d) = \frac{\alpha_d(\alpha_0 - \alpha)}{\alpha_0^2(\alpha_0 + 1)} = \frac{\alpha'_d(1 - \alpha'_d)}{\alpha_0 + 1}$$

$$Cov(p_d, p_f) = \frac{-\alpha_d \alpha_f}{\alpha_0^2(\alpha_0 + 1)} \tag{3}$$

[1] $\mathbb{C}(n) = \mathbb{C}(1)$; n = sum of the multinomials

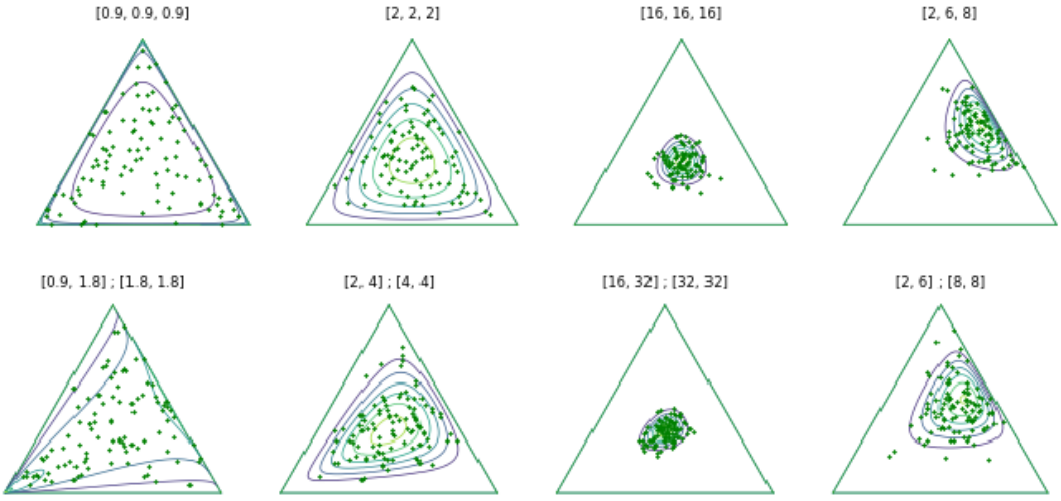

FIGURE 1. Peak of the Dirichlet and generalized Dirichlet distribution at different locations for four different sets of parameter values. First row shows the Dirichlet distribution in the simplex for different values α and second row is the Generalized Dirichlet distribution for different values of α and β.

It should be noted that, small values of the concentration parameter α_0 favors the extreme values of the density function and as a result, data are distributed all over the simplex data and it is more compact at the corner of the simplex. The shape parameter $\boldsymbol{\alpha}$ makes it possible to model data in linear, convex and concave hulls [5]. Figure 1 shows the flexibility of the distribution by changing the parameters. α_0 controls the peak of the distribution and α_d determines the location of the peak. If the expected values of the parameters are equal then data are distributed uniformly over the simplex. The higher the parameter value, more confident we are about that parameter and hence density values are more peaked on that side.

B. GENERALIZED DIRICHLET DISTRIBUTION

From (3), we see that any two random variables following Dirichlet distribution are negatively correlated. If the variables are positively correlated, then Dirichlet prior is not a proper choice. A modification in such cases is the generalized Dirichlet (GD) distribution which entertains both negatively and positively correlated random variables [30]. In a D dimensional closed simplex, generalized Dirichlet distribution with parameter vector $\boldsymbol{\theta} = (\alpha_1, \beta_1, \alpha_2, \beta_2, \ldots, \alpha_D, \beta_D)$ is defined as,

$$P(p|\boldsymbol{\theta}) = \prod_{d=1}^{D} \frac{\Gamma(\alpha_d + \beta_d)}{\Gamma\alpha_d \Gamma\beta_d} p_d^{\alpha_d - 1} \Big(1 - \sum_{l=1}^{d} p_l\Big)^{\gamma_d} \tag{4}$$

Here, $\sum_{d=1}^{D} p_d < 1$, and $0 < p_d < 1$ for $d = 1, 2, \ldots, D$ where $\alpha_d > 0, \beta_d > 0$ and $\gamma_d = \beta_d - \alpha_{d+1} - \beta_{d+1}$, $\gamma_D = \beta_D - 1$ for $d = 1, 2, \ldots, D$. GD becomes Dirichlet distribution when $\beta_d = \alpha_{d+1} + \beta_{d+1}$. If a vector $p \sim GD(\alpha_1, \beta_1, \ldots, \alpha_D, \beta_D)$, then it can be transformed to follow independent Beta distributions for each dimension using the following transformation proposed by [31].

$$z1 = p1 \tag{5}$$

$$z_d = \frac{p_d}{1 - \sum_{j=1}^{d-1} p_j} \tag{6}$$

$$p_d = z_d(1 - p_1 - p_2 -, \ldots, -p_{d-1}) = z_d \prod_{j=1}^{d-1}(1 - z_j) \tag{7}$$

It is evident that generalized Dirichlet distribution has 2D parameters. Unlike Dirichlet distribution where the expectation is fixed, in GD distribution, the expectation of each dimension d continues to evolve over the dimension $d - 1$.

$$E[p_d] = \frac{\alpha_d}{\alpha_d + \beta_d} \prod_{j=1}^{d-1} \frac{\beta_j}{\alpha_j + \beta_j} \tag{8}$$

$$Cov(p_d, p_f) = E(p_f)\Big(\frac{\alpha_d}{\alpha_d + \beta_d + 1} \prod_{j=1}^{d-1} \frac{\beta_j + 1}{\alpha_j + \beta_j + 1}\Big) \tag{9}$$

where, $d, f = 1, 2, \ldots, D$. Flexible covariance structure of GD distribution enables it to have different degrees of belief on random variables while keeping the same expectation [30]. From Fig. 1, it is evident that for Dirichlet distribution, symmetrically distributed random variables are less concentrated at the center (for example, $\alpha = [2, 2, 2]$) than the random variables following generalized Dirichlet distribution which are more concentrated at the center asymmetrically ($\alpha = [2, 4]; \beta = [4, 4]$). It can be shown that generalized Dirichlet distribution reduces to Dirichlet distribution when $\beta_d = \alpha_{d+1} + \beta_{d+1}$ (see [12] for details). If the expectation is varied and for example when $\alpha = [2, 6]; \beta = [6, 8]$ in Fig. 1, generalized Dirichlet distribution captures the variation of the data more flexibly.

C. BETA-LIOUVILLE DISTRIBUTION

While generalized Dirichlet distribution is more flexible than Dirichlet distribution, it requires twice the number of parameters. An efficient replacement for Dirichlet and generalized Dirichlet distribution is the Beta-Liouville distribution

which overcomes the limitations of Dirichlet distribution and requires lesser parameters to estimate than generalized Dirichlet distribution [32]. This distribution is a a special case of Liouville family of distributions. Vector, $p = \{p_1, p_2, \ldots, p_D\}$ will follow a Liouville distribution if and only if $p \overset{d}{=} uq$ where $q = \{q_1, q_2, \ldots, q_D\} = \{\frac{p_1}{\sum p}, \frac{p_2}{\sum p}, \ldots, \frac{p_D}{\sum p}\} \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_D)$ where $\sum p$ is the normalizing constant of vector sum and $u = \sum_{d=1}^{D} p_d$ is an independent random variable. The joint probability density function of this distribution is given by [32],

$$P(p|\alpha_1, \ldots, \alpha_D; \alpha, \beta) = \frac{\Gamma\alpha_0}{B(\alpha, \beta)} \prod_{d=1}^{D} \frac{p_d^{\alpha_d - 1}}{\Gamma\alpha_d} (\sum_{d=1}^{D} p_d)^{\alpha_d - \alpha_0} \times (1 - \sum_{d=1}^{D} p_d)^{\beta - 1} \quad (10)$$

It is evident that the Beta-Liouville distribution has 2 additional parameters than Dirichlet distribution. The mean, variance and covariance of the Beta-Liouville distribution are expressed as follows [32].

$$E[p_d] = \frac{\alpha}{\alpha + \beta} \frac{\alpha_d}{\alpha_0} \quad (11)$$

$$Var(p_d) = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)} \frac{\alpha_d(\alpha_d + 1)}{\alpha_0(\alpha_0 + 1)} - \frac{\alpha^2}{(\alpha + \beta)^2} \frac{\alpha_d^2}{\alpha_0^2} \quad (12)$$

$$Cov(p_d, p_f) = \frac{\alpha_d \alpha_f}{\alpha_0} \Big[\frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)(\alpha_0 + 1)} - \frac{\alpha^2}{(\alpha + \beta)^2 \alpha_0} \Big]; \; d \neq f \quad (13)$$

From (13), we see that Beta-Liouville distribution has more generalized covariance structure compared to negative covariance of Dirichlet distribution. In addition, two random variables with same expectation can have different variances. Such properties of Beta-Liouville distribution makes it more flexible to estimate density of proportional data.

D. PARAMETER ESTIMATION

The concentration parameter $\boldsymbol{\alpha}$ can be determined from the observed proportional data $\boldsymbol{D}_{obs}$ which consists of N observation and each observation is a D dimensional proportional vector. The the joint probability function of the whole dataset can be computed as follows,

$$p(\boldsymbol{D}_{obs}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} p(P_i|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \frac{\Gamma(\sum_d \alpha_d)}{\prod_d \Gamma\alpha_d} \prod_d p_{i,d}^{\alpha_d - 1} \quad (14)$$

In order to maximize (14), we need to take the gradient and set it to zero. It is cumbersome to apply chain rule with the product terms in (14). Therefore, we take maximum likelihood estimation (MLE) approach. Since the distributions discussed above are from exponential family, taking the logarithm will turn it into a convex optimization problem [33] and thus a line search algorithm such as Newton-Raphson method or fixed point iteration method can be applied [34]–[36].

$$\log(p(\boldsymbol{D}_{obs}|\boldsymbol{\alpha})) = N \log \Gamma \sum_d \alpha_d - N \sum_d \log \Gamma\alpha_d + N \sum_d (\alpha_d - 1) \log \bar{p_d} \quad (15)$$

The derivative for one α_d is,

$$g_d = N \, \psi(\sum_d \alpha_d) - N \, \psi(\alpha_d) + N \log \bar{p_d} \quad (16)$$

where $\psi(x) = \frac{d \log \Gamma x}{dx}$ is the digamma function. The gradient, $\mathbf{g}$ for the dataset is $D \times 1$ and can be written as follows,

$$\nabla \log(p(\boldsymbol{D}_{obs}|\boldsymbol{\alpha})) = N \begin{pmatrix} \psi(\sum_d \alpha_d) - \psi(\alpha_1) + \log \bar{p}_1 \\ \psi(\sum_d \alpha_d) - \psi(\alpha_2) + \log \bar{p}_2 \\ \vdots \\ \psi(\sum_d \alpha_d) - \psi(\alpha_D) + \log \bar{p}_D \end{pmatrix} \quad (17)$$

In exponential family of distribution, when the gradient is set to zero, the observed and sufficient statistics becomes equal and as since Dirichlet distribution is from the exponential family, it is possible to formulate an equation and solve it as a fixed point iteration problem to determine the concentration parameters α (see [34] for details). For a vector, this can be expressed as follows-

$$\mathbb{E}[\log p_d] = \psi(\alpha_d) - \psi(\sum_d \alpha_k) \quad (18)$$

$$\psi(\alpha_d^{new}) = \psi(\sum_d \alpha_d^{old}) + \log \bar{p}_k \quad (19)$$

Fixed point iteration method converges only when $|g| < 1$ and is linearly convergent meaning that decreasing error in each step is roughly proportional to previous step. In contrast, Newton-Raphson method solves has quadratic convergence rate and guarantees to converge given that the initial guess is close to final estimate. The Hessian of the log-likelihood function is,

$$H = \nabla^2 \log(p(\boldsymbol{D}_{obs}|\boldsymbol{\alpha})) = \begin{pmatrix} \frac{\partial l^2}{\partial \alpha_1{}^2} & \frac{\partial l^2}{\partial \alpha_1 \alpha_2} & \cdots & \frac{\partial l^2}{\partial \alpha_1 \alpha_d} \\ \frac{\partial l^2}{\partial \alpha_2 \alpha_1} & \frac{\partial l^2}{\partial \alpha_2{}^2} & \cdots & \frac{\partial l^2}{\partial \alpha_2 \alpha_d} \\ & \vdots \ddots & & \\ \frac{\partial l^2}{\partial \alpha_d \alpha_1} & \frac{\partial l^2}{\partial \alpha_d \alpha_2} & \cdots & \frac{\partial l^2}{\partial \alpha_d{}^2} \end{pmatrix} = B + 1_d 1_d^{\mathrm{T}} b \quad (20)$$

where $B \hat{=} \text{diag}: \mathbb{R}^D \to \mathbb{R}^{\mathrm{DxD}} : -N \, \text{diag}(\psi'(\alpha_1), \ldots, \psi'(\alpha_D))$ and $b = N\psi'(\sum_d \alpha_d)$; $\psi'(x) = \frac{\mathrm{d}\psi(x)}{\mathrm{d}x}$ is the trigamma

function. For Newtons algorithm, the Hessian needs to be inverted and [37] provided the following inversion technique using Sherman-Liberman formula-

$$H^{-1} = B^{-1} - \frac{B^{-1}1_D 1_D^T B^{-1}}{b^{-1} + 1_D^T B^{-1} 1_D} \tag{21}$$

Therefore, update for the Newton's algorithm becomes,

$$\boldsymbol{\alpha}^{new} = \boldsymbol{\alpha}^{old} - H^{-1} g \tag{22}$$

As discussed, it is important to estimate the initial values of the parameters more accurately rather than taking random initial guess so that (22) converges to global optima. There are some propositions for the initial estimation of these parameters. Method of moments technique provides good estimate of the initial guess of the parameters. The first and second moment of the data can be calculated from the moment generating function. The moment generating function of a vector X of random variable x is given by $\mathbb{E}(e^{tX})$ and is defined by $\mathbb{M}_X(t)$. With the utilization of Taylor series expansion solving the general moment equation for Dirichlet distribution results in the first and second moments of the Dirichlet distribution presented as follows-

$$\mathbb{E}(X) = \frac{\alpha_d}{\sum_d \alpha_d} \tag{23}$$

$$\mathbb{E}(X^2) = \frac{\alpha_d(\alpha_d + 1)}{\sum_d \alpha_d(\sum_d \alpha_d + 1)} \tag{24}$$

Solving the above equations, we get the values of the parameters $\boldsymbol{\alpha}$ which can be used as an initial guess for the Newton's algorithm.

$$\alpha_d = \mathbb{E}[p_d]\frac{\mathbb{E}[p_d] - \mathbb{E}[p_d^2]}{\mathbb{E}[p_d^2] - \mathbb{E}[p_d]^2} \tag{25}$$

Other techniques such as Expectation Maximization and Expectation Maximization gradient algorithm can also be employed to deter the parameters of the Dirichlet distribution [38].

IV. CLASSIFIER-SUPPORT VECTOR MACHINE

SVM is a well known and common choice for the the supervised machine learning problem. Empirically it has shown good generalization performance in different fields of research and applications [39]–[41]. The aim of using this classifier is to find the support vectors that maximizes the margin between class labels where number of support vectors is proportional to generalization error [42]. Considering the primal representation of the optimization problem, we have

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_i^N \xi_i \\ &\text{subject to, } y^{(i)}(w^T\phi(p_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N \\ &\qquad \xi_i \geq 0, \quad i = 1, \ldots, N \end{aligned} \tag{26}$$

Assume the dataset $\boldsymbol{D}_{obs} = \{(p_i, y_i)\}_i^N$ where N is the number of images and each image is represented by a $L1$-normalized histogram (p_i) and the corresponding label y_i. The objective is to determine the infinite number of linear classifiers that maximizes the geometric margin between the classes with the lowest generalization error. In case of non-separable data, we look into higher dimensional space to find the appropriate hyperplane that maximizes the geometric margin and minimizes the misclassification error through some feature mapping technique. To control the trade off between the large margin and error rate, the hyperparameter C is incorporated.

The above is a convex quadratic optimization problem with linear constraints. Solving this problem will result in the maximum geometric margin between classes. Here, $\phi(p_i)$ is the embedding or feature mapping function from the input space, χ to the feature space, $\mathcal{H}$. If no extra features are extracted from the data then this function represents the original input data known as the attributes and the kernel, K which is the inner product between each datapoint become $\langle p_i, p_j\rangle$ instead of $\langle\phi(p_i), \phi(p_j)\rangle$. For non-linearly separated data, slack variable ξ_i are introduced in the objective function and the constraints are modified accordingly. C is a hyperparameter that regularizes our objective function for misclassification. $\sum_i^N \xi_i$ is the upper bound of the generalization error. For hard margin classifier C is set to high value to lower the misclassification error and for soft margin classifier C is set to low values to provide flexibility at boundary region for some data to be miss-classified.

Solving the dual problem is computationally convenient for large datasets. Relaxing the constraints with the help of Lagrange multipliers, dual solution becomes,

$$\begin{aligned} &\max_{\gamma} \sum_i^N \gamma_i - \frac{1}{2}\sum_i^N\sum_j^N \gamma_i\gamma_j y^{(i)}y^{(j)}\langle\phi(p_i), \phi(p_j)\rangle \\ &\text{subject to: } 0 \leq \gamma_i \leq C, \ \sum_i \gamma_i y^{(i)} = 0 \\ &\text{where } i = 1, \ldots N \quad \forall\, \alpha_i, y^{(i)} \end{aligned} \tag{27}$$

Only the support vectors have $\boldsymbol{\gamma}$ values elsewhere it is zero. Getting the support vectors, the decision function classifies the data by comparing the kernel with the support vectors. The decision function of the support vector machine becomes,

$$f(p) = \sum_i^n \gamma_i y^{(i)}\langle\phi(p_i), \phi(p)\rangle \tag{28}$$

V. FEATURE MAPPING: DIRICHLET SVM, GENERALIZED DIRICHLET SVM, BETA-LIOUVILLE SVM

In this section we focus on the primal and dual form of the optimization problem in (26) and (27) to modify the feature mapping function $\phi(p)$. As discussed, optimum performance of SVM depends on the choice of the kernel function and there is no structured procedure to select the kernel function or feature mapping [43]. One of the advantages of embedding input vectors into the feature space is providing flexibility in choosing the mapping function $\phi(p)$ depending on the structure of the data. Taking the advantage of Dirichlet, generalized Dirichlet and Beta-Liouville distributions

for proportional data modelling, a new feature map can be constructed as follows,

$$\phi_j(p_i)=\begin{cases} p_{ij} & j=1,\ldots,D \\ \frac{\Gamma\left(\sum_d \alpha_d\right)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D} p_{id}^{\alpha_d-1} & j=D+1 \\ \prod_{d=1}^{D}\frac{\Gamma(\alpha_d+\beta_d)}{\Gamma\alpha_d\Gamma\beta_d}p_{id}^{\alpha_d-1} & \\ \times\left(1-\sum_{l=1}^{d} p_{il}\right)^{\gamma_d} & j=D+1 \\ \frac{\Gamma\alpha_0}{B(\alpha,\beta)}\prod_{d=1}^{D}\frac{p_{id}^{\alpha_d-1}}{\Gamma\alpha_d}(\sum_{d=1}^{D}p_{id})^{\alpha_d-\alpha_0} & \\ \times(1-\sum_{d=1}^{D} p_{id})^{\beta-1} & j=D+1 \end{cases} \tag{29}$$

To estimate the parameters in (29), a similar technique is followed as described by [5]. Using the kernel trick, the proposed feature mapping technique can be used with the traditional non-linear kernels to map input space into feature space implicitly without knowing about the feature space. The dimension of the input space is increased by 1 by doing the feature mapping mentioned in (29). We can formulate the Dirichlet SVM (DSVM) as follows,

$$\begin{aligned} &\min_{w,b,\xi}\ \frac{1}{2}\sum_{d=1}^{D+1} w_d^2 + C\sum_{d=1}^{D+1}\xi_i \\ &\quad y^{(i)}\Big(\sum_{d=1}^{D} w_d p_{id} + w_{D+1}\frac{\Gamma\left(\sum_d \alpha_d\right)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\prod_{d=1}^{D} p_i d^{\alpha_d-1} + b\Big) \\ &\quad \geq 1-\xi_i, \quad i=1,\ldots,n \\ &\quad p_{iD} = 1-\sum_{d=1}^{D-1} p_d \\ &\quad \xi_i \geq 0, \quad i=1,\ldots,n \end{aligned} \tag{30}$$

In a similar fashion, generalized Dirichlet SVM (GDSVM) and Beta-Liouville SVM (BLSVM) can be formulated. For a new data p', the trained Dirichlet parameter α is used to determine the feature mapping according to (29). The decision function for this new data becomes,

$$f(p') = \sum_{i}^{N}\left(\gamma_i\sum_{d=1}^{D+1} p_{id}p'_d\right) \tag{31}$$

Applying the flexible mapping function $\phi(p)$ in (29) changes the similarity measure and thus enables us to modify the base kernel. Apart from the regular kernels such as RBF, polynomial, sigmoid, χ^2 which are discussed vastly in the literature, we combine our proposed feature mapping technique with the following kernels as well,

- **Bhattacharya Measure**
  Bhattacharya coefficient is a divergence type measure between distributions and defined as [44],

$$B = \sum_{i=1}^{N}\sqrt{p_i q_i} \tag{32}$$

Considering a $D+1$ dimensional vector, it can be geometrically interpreted that the Bhattacharya coefficient measures the cosine of the angle between the vector elements. Since, p_i and q_i represent probability distributions and if they have the similar density function then the coefficient is 1. However, this coefficient can not be used as a metric distance since it violates the axioms of being a distance metric [45]. To make a proper representation of the distance metric, [44] modified the coefficient as $D_{p_i,q_i} = \sqrt{1-B}$. The kernel for this distance with hyperparameter γ,

$$K(p,q) = e^{-\gamma\sqrt{1-B}} \tag{33}$$

- **Generalized Histogram Intersection**
  Histogram intersection kernel is a positive definite kernel and satisfies Mercer's condition to be used in SVM [2], [46]. Global or low-level features are commonly used for this, however, use of local features works well with this kernel as well. Given two vectors namely p_i and p_j containing the elements of two normalized histogram, histogram intersection measures the similarity between the them by using (34) [47].

$$K(p,q) = \sum_{i=1}^{N} min[(p_i)^\alpha, (q_i)^\alpha] \tag{34}$$

  Setting $\alpha = 1$ results in histogram intersection kernel.
- **Jeffrey Divergence**
  KL-divergence is non-symmetric and sensitive to histogram binning [48]. In addition, it is not robust and does not qualify to be used as a metric of the spread since it violates the triangle inequality. In response to this, Jeffrey divergence is empirically derived and it is mostly invariant to noise and histogram binning [49].

$$K(p,q) = \sum_{i=1}^{N}(p_i\log\frac{p_i}{\mu_i} + q_i\log\frac{q_i}{\mu_i}); \quad \mu_i = \frac{p_i+q_i}{2} \tag{35}$$

- **Rational Quadratic**
  From the probabilistic graphical point of view, several squared error kernels are derived and rational quadratic is one of them. This kernel is a scale mixture of different characteristic length scales [50]. This kernel is useful for modelling data which varies in multiple scales.

$$K(p,q) = \left(1+\frac{\sum_i^N ||p_i-q_i||^2}{2\alpha l^2}\right)^{-\alpha} \tag{36}$$

  Here, α is scale mixture parameter and l is the scale length.
- **Inverse Multiquadratic**
  Inverse multiquadratic function is a member of generalized multiquadratic (GMQ) family of radial basis functions defined by $K(p,q) = (c^2 + (\epsilon r)^2)^\beta$ [51] where ϵ is the shape parameter and parameter β determines the positive definiteness of the function [52].

Unlike multiquadratic kernel, inverse multiquadratic is a positive definite [53]. Setting $\beta = \frac{1}{2}$, we get the following expression for this kernel-

$$K(p, q) = \frac{1}{\sqrt{\sum_i^N |p_i - q_i|^2 + c^2}} \tag{37}$$

- **ANOVA**
  ANOVA kernel is one of the examples of convolution kernels [54]. This kernel uses factor d to get higher order interactions of the features that we are interested in and then sum over the terms to get the similarity score.

$$K(p, q) = \sum_i^N e^{-(\sigma(p_i - q_i)^2)^d} \tag{38}$$

- **Generalized T-student Kernel**
  This is a positive semi definite kernel and satisfies the condition of Mercer's theorem [55]. It has similar form to Inverse Multiquadratic kernel.

$$K(p, q) = \sum_i^N \frac{1}{1 + (p_i - q_i)^d} \tag{39}$$

- **MinMax**
  MinMax is a graph kernel proposed by [56] which is similar to Tanimoto kernel when applied to binary dataset. MinMax kernel models count data and thus takes into account the values between 0 and 1. Therefore, this kernel is suitable for proportional data modelling.

$$K(p, q) = \frac{\sum_i^N \min(p_i, q_i)}{\sum_i^N \max(p_i, q_i)} \tag{40}$$

- **Cauchy**
  Derived from the long tail Cauchy distribution, Cauchy kernel puts more weight on interaction of distant non-zero values [57]. [58] applied the Cauchy kernel for sparse coding of natural scenes data.

$$K(p, q) = \sum_i^N \frac{1}{1 + \frac{(p_i - q_i)^2}{s^2}} \tag{41}$$

  Unlike Gaussian kernel, in this kernel moving from the center gives more weight to the features. A combination of these two kernels showed good classification performance on some dataset [57].
- **Cosine Similarity**
  In an inner product space, cosine similarity measures the similarity between the two vectors by calculating the direction of each vector [59]. This is a non-metric measure since it does not satisfy all the conditions to be a metric.

$$K(p, q) = \frac{\langle p_i, q_i \rangle}{||p_i|| ||q_i||} \tag{42}$$

Algorithm 1 Algorithm for DSVM, GDSVM and BLSVM

1. **Input:** Training data,
$\boldsymbol{D}_{obs} = \{(p_1, y_1), (p_2, y_2), \ldots, (p_N, y_N)\}$.
2. **Estimate:** Initial parameters using Method of Moments (MoM) [5].
3. **Update:** Apply Newton Raphson's method until convergence [5].
4. **Compute kernel:**
   - Base kernel: Compute K(**p**, **q**) from (33) to (45) for $\phi_j(p_i)$ in (29) only when $j = 1, 2, \ldots, D$.
   - DSVM: Use first and second form of (29) for $\phi_j(p_i)$ and apply (33) to (45) to compute DSVM kernel, K(**p**, **q**).
   - GDSVM: Use first and third form of (29) for $\phi_j(p_i)$ and apply (33) to (45) to compute GDSVM kernel, K(**p**, **q**).
   - BLSVM: Use first and fourth form of (29) for $\phi_j(p_i)$ and apply (33) to (45) to compute BLSVM kernel, K(**p**, **q**).
5. **Optimization:** Solve the primal problem in (26) or dual problem in (27) to get the support vectors.

- **Tanimoto or Extended Jaccard Similarity**
  A modification in the cosine similarity function results in Tanimoto similarity index [56]. It represents the number of attributes shared by the vectors.

$$K(p, q) = \frac{\langle p_i, q_i \rangle}{\langle p_i, p_i \rangle + \langle q_i, q_i \rangle - \langle p_i, q_i \rangle} \tag{43}$$

  Here, $\langle p_i, q_i \rangle = \sum_{d=1}^D p_{id} \times q_{id}$ and the term $\langle p_i, p_i \rangle = ||p_i||^2$ and $\langle q_i, q_i \rangle = ||q_i||^2$ is the Euclidean norm or the length of the vector. [60] derived the modified Tanimoto coefficient in relation with Cosine similarity as,

$$K(p, q) = \frac{\text{cossim}(p_i, q_j)}{\frac{||p_i||^2 + ||q_i||^2}{||p_i|| ||q_i||} - \text{cossim}(p_i, q_i)} \tag{44}$$

  Here, $\text{cossim}(x_i, y_j)$ is calculated from (42).
- **Sorensen Similarity**
  Similar to cosine similarity Sorensen similarity index is a non-metric measure as it does not satisfy all the axioms of being a metric. This measure is more appropriate in retaining the sensitivity of the heterogeneous data than Euclidean distance and in image segmentation and lexical association [61], [62].

$$K(p, q) = \sum_i^N \frac{2 p_i q_i}{p_i^2 + q_i^2} \tag{45}$$

Algorithm 1 shows the steps for the Dirichlet SVM, generalized Dirichlet SVM and Beta Liouville SVM using (29).

VI. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed feature mapping technique for natural scene classification, satellite image

TABLE 1. Scene recognition performance results for baseline kernels and our proposed kernels.

Kernel	Baseline SVM	DSVM	GDSVM	BLSVM
Linear	0.72000	0.72000	0.70667	**0.74000S**
Polynomial	0.76000	**0.77333**	0.76000	0.74677
Sigmoid	0.70677	0.72000	0.72000	**0.73333**
RBF	0.70677	0.72000	0.71333	**0.74677**
Exponential	0.74667	0.74667	0.74667	**0.79333**
Tanimoto	0.74000	0.74000	0.74000	**0.76000**
MinMax	0.76667	0.76000	0.76000	0.76667
Bhattacharya	0.74000	0.74667	0.73333	**0.76000**
Cosine Similarity	0.72667	0.71333	0.71333	**0.73333**
Rational Quadratic	0.75333	**0.76000**	**0.76000**	0.72667
Inverse Multiquadratic	0.77333	0.77333	**0.78000**	0.74000
Cauchy	0.71333	0.72000	0.72000	**0.75333**
Tstudent	0.75333	**0.76000**	**0.76000**	0.72667
ANOVA	0.72667	0.71333	0.72000	**0.74667**
Sorensen Similarity	0.72667	0.72667	0.72667	0.71333
Additive χ^2	0.76667	**0.77333**	0.76667	0.76000
Histogram Intersection	0.76000	0.76000	0.76000	0.74667

TABLE 2. Satellite image classification performance results for baseline kernels and our proposed kernels.

Kernel	Baseline SVM	DSVM	GDSVM	BLSVM
Linear	0.86364	0.85577	0.87000	**0.90196**
Polynomial	0.85454	0.86486	0.85454	**0.89189**
Sigmoid	0.86274	0.86274	0.86364	**0.89215**
RBF	0.87272	0.87272	0.87272	**0.89215**
Exponential	0.88073	0.88073	**0.88991**	0.88182
Tanimoto	0.90566	0.90566	0.90566	**0.90999**
MinMax	0.88462	0.89423	0.88462	**0.90197**
Sorensen Similarity	0.87273	0.86363	0.88182	**0.89216**
Bhattacharya	0.90196	0.89215	0.90196	**0.88991**
Cosine Similarity	0.86363	0.87000	0.87273	**0.90196**
Rational Quadratic	0.88181	0.87272	0.86363	**0.88235**
Inverse Multiquadratic	0.88000	0.88000	0.88000	**0.88235**
Cauchy	0.88000	0.89000	0.89000	**0.89215**
Tstudent	0.86000	0.87000	0.86000	**0.88235**
ANOVA	0.85294	0.85000	0.86000	**0.87129**
Additive χ^2	0.89215	0.89215	0.89215	**0.90826**
Histogram Intersection	0.90384	0.90384	0.89423	**0.91176**
Jfd	0.89215	0.89215	0.88679	**0.90196**

classification and human action recognition in videos. The dual form of the SVM optimization problem is solved using [63]. For multi-class classification, one-vs-all technique is applied and the tolerance value 10^{-3} is used as stopping criterion and a hard limit on the solver is imposed by setting maximum iterations to 5000. All the models are evaluated using 10 fold cross validation. 9 folds are used for training and the remaining fold for testing the model. Similar to [5], for image classification best score is reported for each kernel and for action recognition, average score with standard deviation are reported for all kernels. For misclassification, the hyperparameter C in the objective function is varied from 1 to 15 in 10 base logarithm scale and best models are found by doing a simple grid search and are reported thereby. For polynomial kernel, degree 3 is considered and for RBF kernel the similarity measurements are scaled down by dividing the length of vocabulary size. In general, BLSVM performs better than DSVM and GDSVM approaches. As mentioned, generalized Dirichlet distribution has twice the number of parameters than Dirichlet distribution and density values are more concentrated around the mean compared to Dirichlet distribution (in Fig.1). Since our approach is to perform feature mapping after combining discriminative features with distribution based features, we assume that the feature pair similarity values in similarity matrix for generalized Dirichlet distribution are hard to separate after solving the dual form

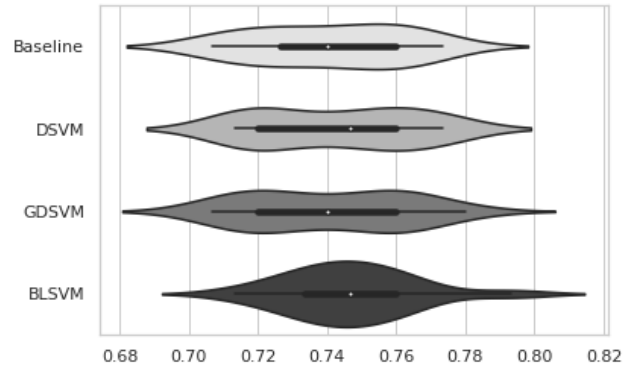

FIGURE 2. Violin plots of experimental results for 15 scenes dataset.

SVM in (27). Therefore, Beta-Liouville distribution is proved to be a better generalization of Dirichlet distribution in our proposed approach.

A. 15 SCENES DATASET CLASSIFICATION

Scene recognition is very essential for reasoning in navigation and recognition tasks. Specially in terms of robotics and automation it is significant to enhance machine's visual understandings [64]. 15 scene dataset consists of 15 different scene categories. First 13 categories were collected combinedly by [65] and [66]. For our experiment, from each category 100 images were selected totalling to 1500 images. Local features are extracted using Scale Invariant Feature Transform (SIFT) [67] algorithm as it is invariant to scale and rotation. In our experiment, we calculate dense SIFT [65] for speed using [68]. Descriptors are computed for densely sampled keypoints with similar size and orientation. Each images is converted to gray-scale and for each pixel descriptors are computed over a patch of 16×16 pixels. The extracted features are quantized into a vocabulary size of 200. Table 2 shows the best results for the baseline SVM, DSVM, GDSVM and BLSVM.

We can see that, with our proposed feature mapping technique accuracy score for the classification task has significantly improved. The reason is because of increased separability among the support vectors from each image category. In the case of linear feature map, BLSVM shows a 2% improvement in accuracy score. Non-linear kernels with BLSVM performs better than DSVM and GDSVM. We conduct statistical hypothesis testing (t-text) to investigate the scores of each approaches. Results of DSVM and BLSVM are statistically significant (p-value $< 5\%$). However, performance difference of baseline score and GDSVM is not statistically significant for this dataset (p-value of 0.40). Mean average accuracy of BLSVM is 74.67% compared to DSVM's 74.27%. Thus, BLSVM is the preferred method for this dataset which requires only 2 more parameters to learn than DSVM. Such improvement is perhaps because of Beta-Liouville distribution's better generalization capabilities shown in (11)-(13) to capture data distribution with less number of parameters. Fig.2 shows the probability distribution of the experimental results presented in Table 2. It is evident that Beta-Liouville distribution based feature mapping can be used more confidently with traditional kernels functions. The wider region around the mean in the violin plot represents a higher probability of getting consistent average result.

FIGURE 3. Sample image from 15 different categories: 1. bedroom, 2. sea coast, 3. field, 4. forest, 5. highways, 6. house, 7. industrial, 8. kitchen, 9. living room, 10. mountain, 11. stadium, 12. store, 13. street, 14. sky scrappers, 15. ocean underwater.

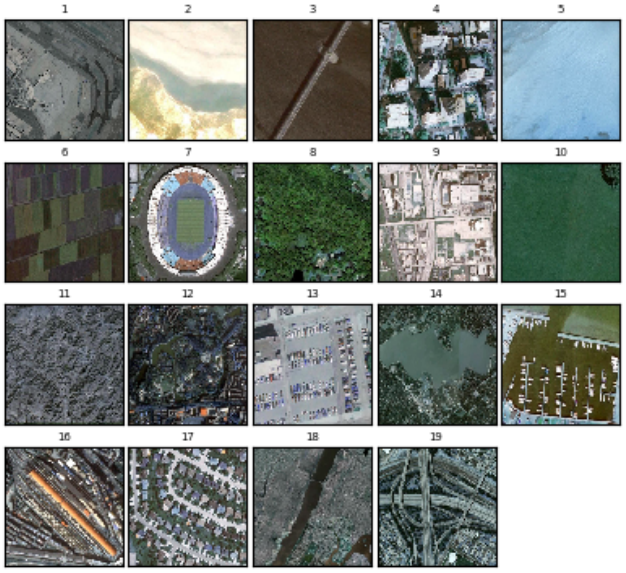

FIGURE 4. Sample satellite image from 19 different categories: 1. airport, 2. sea beach, 3. bridge, 4. commercial area, 5. desert, 6. farmland, 7. stadium, 8. forest, 9. industrial area, 10. meadow, 11. mountain, 12. park, 13. parking, 14. pond, 15. port, 16. railway station, 17. residential area, 18. river, 19. viaduct.

B. SATELLITE IMAGE CLASSIFICATION

This dataset has 19 categories of google satellite images collected from http://www.escience.cn/people/yangwen/WHU-RS19.html. Each category has 50 images and the resolution of each image is 600×600. The challenges in classifying high resolution satellite image data is that the dominance of structures and objects in the image leads to misclassification [69]. For feature extraction, we use the same configuration as described in previous section.

FIGURE 5. Sample frame from each categories performed by one person. Each frame is resized to 160 × 120.

TABLE 3. 10 fold cross validation results of action recognition from videos.

Kernel	Baseline SVM	DSVM	GDSVM	BLSVM
Linear	0.90401 ±0.047	0.89886 ±0.048	0.90401 ±0.047	**0.91167 ±0.056**
Polynomial	0.82869 ± 0.045	0.92323 ± 0.042	0.84132 ± 0.045	**0.93185 ± 0.043**
Sigmoid	0.89171 ± 0.050	0.92046 ± 0.047	0.90045 ± 0.059	**0.93185 ±0.049**
RBF	0.90197 ±0.050	0.92319 ±0.042	0.89449 ±0.056	0.93185 ±**0.043**
Exponential	0.91161 ±0.054	0.91399 ±0.051	0.90962 ±0.050	0.91677 ±**0.047**
Tanimoto	0.90923 ±0.048	0.92040 ±0.042	0.90923 ±0.048	**0.92868 ±0.046**
MinMax	0.92034 ± 0.051	**0.94104 ±0.045**	0.93933 ±0.059	0.93661 ±0.031
Sorensen Similarity	0.89934 ±0.064	0.92041 ±0.042	0.90214 ±0.064	**0.92590 ±0.045**
Bhattacharya	0.90634 ± 0.040	0.90634 ± 0.040	0.90395 ± 0.044	**0.92403 ± 0.044**
Cosine Similarity	0.88701 ± 0.064	**0.89528 ± 0.053**	0.88939 ± 0.063	0.89296 ± 0.049
Rational Quadratic	0.89897 ± 0.066	0.92046 ± 0.047	0.89858 ± 0.061	**0.92629 ± 0.040**
Inverse Multiquadratic	0.90969 ± 0.044	0.92046 ± 0.047	0.90969 ± 0.044	**0.92392 ± 0.047**
Cauchy	0.89212 ± 0.053	0.91008 ± 0.048	0.89212 ± 0.053	**0.91473 ± 0.056**
ANOVA	0.89767 ± 0.064	0.89767 ± 0.064	**0.90481 ± 0.062**	0.90124 ± 0.056
Additive χ^2	0.91989 ± 0.052	**0.92312 ± 0.051**	0.92267 ± 0.049	0.92205 ±0.048
Histogram Intersection	0.92001 ± 0.065	**0.93826 ± 0.045**	0.92278 ± 0.061	0.93383 ± 0.031
Jfd	0.90963 ± 0.061	**0.92312 ± 0.051**	0.90481 ± 0.057	0.91689 ± 0.051

For all the kernel, BLSVM outperforms baseline SVM, DSVM and GDSVM except for the exponential kernel where generalized Dirichlet SVM achieves higher accuracy of 88.991%(Table 2). Considering the core form SVM, BLSVM gives highest accuracy of 90.196% whereas linear SVM achieves 86.364% accuracy. Smaller p-value (less than 0.005) from Student's t-test confirms that performance results obtained from BLSVM are statistically significant and thus we reject the null hypothesis of being equal averages with other approaches. Fig.6 shows the distribution of accuracy score for BLSVM has less variance than other approaches which guarantees that it can be used confidently with selected kernels for this dataset.

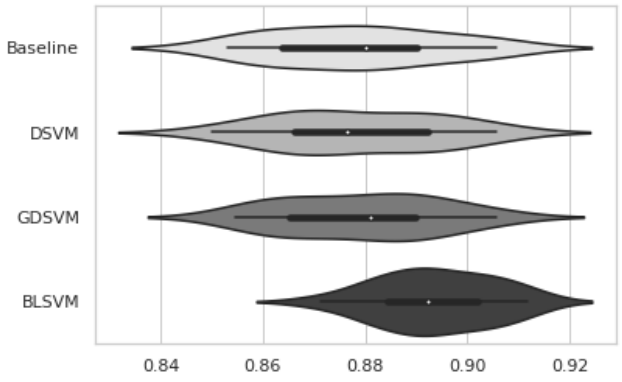

FIGURE 6. Violin plots of experimental results for Satellite image classification dataset.

C. HUMAN ACTION RECOGNITION

Recognizing human action in videos is an interesting learning task for surveillance and navigation tasks. For the purpose of evaluation of our model for videos, we choose KTH-human action recognition data introduced by [7]. This dataset contains 6 categories each having 100 videos with 4 different scenarios and each action is performed by 25 different persons with different variations e.g. different color of clothing, different motion of the person, camera angle, zooming, zittering etc. In total, there are 2391 sequences in this dataset. We are interested in dense features as it is more accurate than sparse features. Thus, we use dense optical flow algorithm proposed by [70]. Open source computer vision library such as [71] is used with default values to extract features with the codebook size of 500. Each frame is resized to 160 × 120 and further downsampled to 16 × 12 by taking the pixel values of the positions which are divisible by 10. L_2 normalization is used for feature invariance. To create Dirichlet, generalized

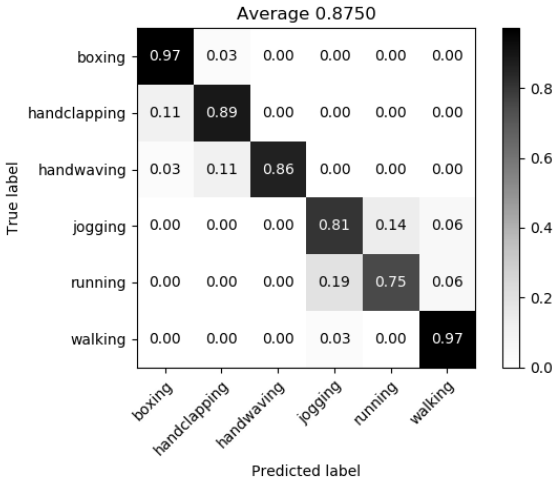

FIGURE 7. Confusion matrix for human action recognition in videos.

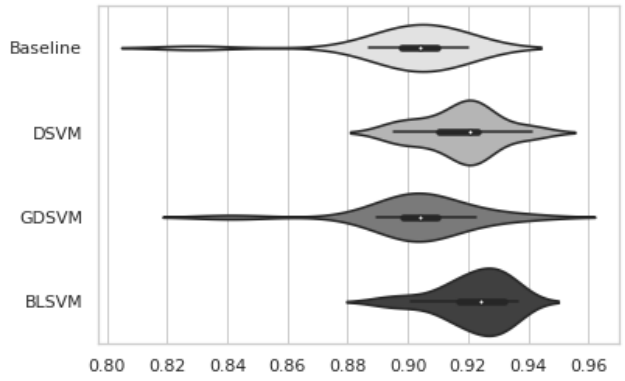

FIGURE 8. Violin plots of experimental results for human action recognition dataset.

Dirichlet and Beta-Liouville SVM, the whole dataset is normalized as proposed in [5]. For 10 fold cross validation, mean accuracy with standard deviation are reported in Table 3. Total 384 videos are used training and 216 videos are used for testing. In the test data, each class has 36 videos.

From Table 3, highest average accuracy of baseline SVM is 92.034% for MinMax kernel which is increased to 94.104% when we combine MinMax kernel with Dirichlet feature mapping function. Fig. 7 shows the confusion matrix for DSVM MinMax kernel which achieves 87.50% accuracy for the test data compared to base MinMax kernel's score of 86.11%. Significance testing using Student's t-distribution shows that DSVM and BLSVM is statistically significant (p-values between 0.0009 to 0.007). GDSVM score is not significantly different than baseline SVM (p-value of 0.7). Heavy right tail of BLSVM's performance distribution in Fig.8 signifies that there is a greater probability of getting a higher accuracy score than DSVM.

VII. CONCLUSION

This paper shows a novel feature mapping technique for proportional data based on Dirichlet, generalized Dirichlet and Beta-Liouville distributions which shows good accuracy in classifying images and videos. Such data are prevalent in data mining, image processing and pattern recognition problems which motivated us to exploit the statistical representation of the data in order to enhance the discriminative power of the traditional SVM kernels. In particular, we have introduced three feature mapping functions for proportional data to be used in SVM learning algorithm. Our experiments show good performance of the proposed technique in classifying natural and satellite images and also in classifying human action recognition in videos. The results also show that either of the proposed distribution based feature mapping functions increases the accuracy of the corresponding SVM kernel.

REFERENCES

[1] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[2] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Proc. Int. Conf. Image Process.*, vol. 3, 2003, p. III-513.
[3] J. P. Singh and N. Bouguila, "Proportional data clustering using K-means algorithm: A comparison of different distances," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2017, pp. 1048–1052.
[4] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, Jun. 1998.
[5] A. Nedaie and A. A. Najafi, "Support vector machine with Dirichlet feature mapping," *Neural Netw.*, vol. 98, pp. 87–101, Feb. 2018.
[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[7] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, pp. 32–36.
[8] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[9] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1996–2003.
[10] T. Kobayashi, "Dirichlet-based histogram feature transform for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3278–3285.
[11] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1055–1064, Sep. 1999.
[12] N. Bouguila, "Hybrid generative/discriminative approaches for proportional data modeling and classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 12, pp. 2184–2202, Dec. 2012.
[13] S. Bourouis, A. Zaguia, N. Bouguila, and R. Alroobaea, "Deriving probabilistic SVM kernels from flexible statistical mixture models and its application to retinal images classification," *IEEE Access*, vol. 7, pp. 1107–1117, 2019.
[14] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, Dec. 2013.
[15] N. Bouguila and O. Amayri, "A discrete mixture-based kernel for SVMs: Application to spam and image categorization," *Inf. Process. Manage.*, vol. 45, no. 6, pp. 631–642, Nov. 2009.
[16] A. Sefidpour and N. Bouguila, "Spatial color image segmentation based on finite non-Gaussian mixture models," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8993–9001, Aug. 2012.
[17] N. Bouguila, "Deriving kernels from generalized Dirichlet mixture models and applications," *Inf. Process. Manage.*, vol. 49, no. 1, pp. 123–137, Jan. 2013.
[18] G. Lebanon, "Metric learning for text documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 497–508, Apr. 2006.
[19] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1932–1939.

[20] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2046–2053.
[21] O. Pele and M. Werman, "The quadratic-chi histogram distance family," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 749–762.
[22] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.
[23] H. Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[24] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1447–1454.
[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[26] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.
[27] A. Nedaie and A. A. Najafi, "Polar support vector machine: Single and multiple outputs," *Neurocomputing*, vol. 171, pp. 118–126, Jan. 2016.
[28] Z. Li, X. Yang, W. Gu, and H. Zhang, "Kernel-improved support vector machine for semanteme data," *Appl. Math. Comput.*, vol. 219, no. 17, pp. 8876–8880, May 2013.
[29] K. W. Ng, G.-L. Tian, and M.-L. Tang, *Dirichlet and Related Distributions: Theory, Methods and Applications* (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 2011.
[30] T.-T. Wong, "Generalized Dirichlet distribution in Bayesian analysis," *Appl. Math. Comput.*, vol. 97, nos. 2–3, pp. 165–181, Dec. 1998.
[31] R. J. Connor and J. E. Mosimann, "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *J. Amer. Stat. Assoc.*, vol. 64, no. 325, pp. 194–206, Mar. 1969.
[32] N. Bouguila, "Bayesian hybrid generative discriminative learning based on finite liouville mixture models," *Pattern Recognit.*, vol. 44, no. 6, pp. 1183–1200, Jun. 2011.
[33] J. Huang, "Maximum likelihood estimation of Dirichlet distribution parameters," CMU Tech. Rep., 2005.
[34] T. Minka, "Estimating a Dirichlet distribution," MIT, Cambridge, MA, USA, Tech. Rep., 2000.
[35] N. Wicker, J. Muller, R. K. R. Kalathur, and O. Poch, "A maximum likelihood approximation method for Dirichlet's parameter estimation," *Comput. Statist. Data Anal.*, vol. 52, no. 3, pp. 1315–1322, Jan. 2008.
[36] M. Sklar, "Fast MLE computation for the Dirichlet multinomial," 2014, *arXiv:1405.0099*. [Online]. Available: http://arxiv.org/abs/1405.0099
[37] K. S. Miller, *Some Eclectic Matrix Theory*. Melbourne, FL, USA: Krieger, 1987.
[38] K. W. Ng, G.-L. Tian, and M.-L. Tang, *Dirichlet and Related Distributions: Theory, Methods and Applications*, vol. 888. Hoboken, NJ, USA: Wiley, 2011.
[39] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.
[40] X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based component classifiers," *Eng. Appl. Artif. Intell.*, vol. 21, no. 5, pp. 785–795, Aug. 2008.
[41] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retr.*, vol. 12, no. 5, pp. 526–558, Oct. 2009.
[42] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.
[43] T. Bdiri and N. Bouguila, "Bayesian learning of inverted Dirichlet mixtures for SVM kernels generation," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1443–1458, Oct. 2013.
[44] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.
[45] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, K. Fukunaga, Ed., 2nd ed. Boston, MA, USA: Academic, 1990.
[46] S. Boughorbel, J.-P. Tarel, and N. Boujemaa, "Generalized histogram intersection kernel for image recognition," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, Sep. 2005, p. III-161.
[47] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[48] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[49] W. Hu, N. Xie, R. Hu, H. Ling, Q. Chen, S. Yan, and S. Maybank, "Bin ratio-based histogram distances and their application to image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2338–2352, Dec. 2014.
[50] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Berlin, Germany: Springer, 2003, pp. 63–71.
[51] M. Mongillo and I. Institute of Technology, "Choosing basis functions and shape parameters for radial basis function methods," *SIAM Undergraduate Res. Online*, vol. 4, pp. 190–209, 2011.
[52] M. E. Chenoweth and S. A. Sarra, "A numerical study of generalized multiquadric radial basis function interpolation," *SIAM Undergraduate Res. Online*, vol. 2, no. 2, pp. 58–70, 2009.
[53] H. Javaran and N. Khaji, "Inverse multiquadric (IMQ) function as radial basis function for plane dynamic analysis using dual reciprocity boundary element method," in *Proc. 15th World Conf. Earthq. Eng.*, 2012, pp. 24–28.
[54] G. Wahba, *Spline Models for Observational Data*, vol. 59. Philadelphia, PA, USA: SIAM, 1990.
[55] S. Boughorbel, J.-P. Tarel, and F. Fleuret, "Non-mercer kernels for SVM object recognition," in *Proc. BMVC*, 2004, pp. 1–10.
[56] L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi, "Graph kernels for chemical informatics," *Neural Netw.*, vol. 18, no. 8, pp. 1093–1110, Oct. 2005.
[57] J. Basak, "A least square kernel machine with box constraints," *J. Pattern Recognit. Res.*, vol. 5, no. 1, pp. 38–51, 2010.
[58] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, p. 607, 1996.
[59] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
[60] D. C. Anastasiu and G. Karypis, "Efficient identification of tanimoto nearest neighbors," *Int. J. Data Sci. Anal.*, vol. 4, no. 3, pp. 153–172, Nov. 2017.
[61] M. A. Fligner, J. S. Verducci, and P. E. Blower, "A modification of the Jaccard–Tanimoto similarity index for diverse selection of chemical compounds using binary strings," *Technometrics*, vol. 44, no. 2, pp. 110–119, May 2002.
[62] P. Rychlỳ, "A lexicographer-friendly association score," in *Proc. RASLAN*, 2008, pp. 6–9.
[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[64] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "SUN database: Exploring a large collection of scene categories," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 3–22, Aug. 2016.
[65] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 524–531.
[66] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2169–2178.
[67] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[68] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, Mar. 2012.
[69] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.
[70] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scandin. Conf. Image Anal.* Berlin, Germany: Springer, 2003, pp. 363–370.
[71] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision With the OpenCV Library*. Newton, MA, USA: O'Reilly Media, Inc., 2008.

MD. HAFIZUR RAHMAN received the bachelor's degree in industrial engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2017, and the master's degree in quality systems engineering from Concordia University, Montreal, Canada, in 2020. He is currently working as a ML Researcher and Developer at Heyday.ai. His research interests include deep learning, Bayesian deep learning, and graph representation learning and their multidisciplinary applications.

NIZAR BOUGUILA (Senior Member, IEEE) received the Engineering degree in computer science from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D. degrees in computer science from the University of Sherbrooke, Sherbrooke, QC, Canada, in 2002 and 2006, respectively. He is currently a Professor with the Concordia Institute for Information Systems Engineering, Concordia University, Montréal, QC, Canada. His research interests include image processing, machine learning, data mining, 3-D graphics, computer vision, and pattern recognition.