An Effective Feature Selection Method Based on Pair-Wise Feature Proximity for High Dimensional Low Sample Size Data

S L Happy
Department of Electrical Engineering,
Indian Institute of Technology
Kharagpur, India
Email: happy@iitkgp.ac.in

Ramanarayan Mohanty
Advanced Technology
Development Center,
Indian Institute of Technology
Kharagpur, India
Email: ramanarayanmohanty@iitkgp.ac.in

Aurobinda Routray
Department of Electrical Engineering,
Indian Institute of Technology
Kharagpur, India
Email: aroutray@iitkgp.ac.in

***Abstract*—Feature selection has been studied widely in the literature. However, the efficacy of the selection criteria for low sample size applications is neglected in most cases. Most of the existing feature selection criteria are based on the sample similarity. However, the distance measures become insignificant for high dimensional low sample size (HDLSS) data. Moreover, the variance of a feature with a few samples is pointless unless it represents the data distribution efficiently. Instead of looking at the samples in groups, we evaluate their efficiency based on pair-wise fashion. In our investigation, we noticed that considering a pair of samples at a time and selecting the features that bring them closer or put them far away is a better choice for feature selection. Experimental results on benchmark data sets demonstrate the effectiveness of the proposed method with low sample size, which outperforms many other state-of-the-art feature selection methods.**

***Index Terms*—Feature selection, pair-wise feature proximity, high dimensional low sample size data.**

I. Introduction

In this age of information, high dimension data with low sample size are very common in various areas of science [1]. In supervised classification problems, the classification performance is mostly determined by the inherent class information possessed by the features. Hence, it is logical to include more number of features to improve the discriminating ability. In this way, most of the practical machine learning tasks deal with high dimensional data while the number of labeled data are far less than its dimensionality. The feature space for such data is almost empty, and the curse of dimensionality causes the distance measure to become uniform [2]. In addition, the sparsity of labeled instances in high dimensional feature space adversely affects the classification performance as well.

Feature selection is a process of selecting an optimal subset of features from the input feature set based on a selection criterion [3]. Thus, it reduces the data dimensionality by removing redundant features and improves the time and space complexity of the data. In addition, it reduces the risk of over-fitting which is very common in high dimensional data analysis. These algorithms can be categorized as supervised, unsupervised or semi-supervised [4] based on their utilization of label information. Different criteria functions have been proposed in the literature to evaluate the goodness of features, such as mutual information (MutInf) [5], Fisher score (FS) [6], feature selection via concave minimization (FSCM) [7], ReliefF [8], Laplacian score (LS) [9], trace ratio criterion (TRC) [10], spectral feature selection (SPEC) [11], infinite feature selection (IFS) [12] etc. They have demonstrated excellent performance in real-world applications.

The key to obtaining the suitable subset of features depends upon the selection criteria. Algorithms, such as FS and ReliefF, optimize the sample separability, whereas LS preserves sample similarity in the local neighborhood [9]. MutInf considers the mutual information between the distribution of a feature as the selection criterion [5]. SPEC selects the features by analyzing the spectrum of the graph induced from the proximity matrix [11]. However, all these methods evaluate each feature independently and select the top ones based on the utility of features. Such heuristic algorithms neglect the combined performance of multiple features which leads to the selection of a suboptimal subset of features [13]. Thus, it is entirely possible that the performance with two best scoring features may be lower than the performance of any other two features combined. However, finding the global optimal solution is an NP hard problem and very challenging.

Feature selection methods like generalized Fisher score (GFS) [14], TRC and IFS try to globally optimize feature subset to maximize the subset level score. GFS jointly selects features, which maximize the lower bound of traditional Fisher score. Similarly, IFS considers each feature as a node in the affinity graph and assigns a score to each by taking into account all the possible feature subsets as paths on a graph [12]. These methods eliminate the redundant features while considering the combination of features, which gives an advantage over independent feature evaluation methods.

The graph based methods need a suitable number of instances to learn the graph structure, failing which results in inferior performance. Higher accuracy demands sufficient

ISBN 978-0-9928626-7-1 © EURASIP 2017

labeled training data, however, annotating data for real-world applications is an expensive and time consuming job [15]. The methods that compute the features independently are less affected by the low sample size. Nevertheless, the performance of most of the selection criteria decreases with a low number of sample instances.

We address the issues that arise when there is a lack of labeled data samples. When the available samples are less, the only way to select the best features is to assign scores based on how close they are to the samples of the same class while keeping maximum distance from other class samples. We propose a naive way of selecting features which involves the combinational feature selection followed by the heuristic approach of score assignment to each feature. It takes the advantages of selecting a group of features based on the pair-wise proximity in feature values and the lower computational complexity of heuristic search for assigning scores and ranking the features. Instead of using the whole feature vector for distance measurement, we use a subset of the original feature set. Thus, features, responsible for bringing the points of the same class closer while keeping a safe distance from the other class instances, can be found based on the distance between each pair of training instances. Here the basic assumption is that the optimal feature set minimizes the within-class distance, while maximizing the between-class distances for each pair of samples. The proposed pair-wise feature proximity (PWFP) based feature selection method is compared with other literature and evaluated extensively.

A. Notation

In a supervised feature selection scenario, the algorithm is provided with the data and the corresponding class label. Suppose the data set consists of n number of d-dimensional points ($x_i = [x_i^1, x_i^2, ..., x_i^d] \in \mathbb{R}^d$), given by $\{(x_i, y_i)\}_{i=1}^n$. Here $y_i \in \{1, 2, ..., c\}$ represents the class label of the corresponding data. The problem of feature selection aims at finding the feature subset which carries the maximum information to classify the features into accurate classes. Further, we denote the total data matrix as $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d \times n}$, and $f^i = [f^{i1}, f^{i2}, ..., f^{in}]$ represents the ith row of the matrix X. Without loss of generality, we assume that X has been centered with zero mean, i.e., $\sum_{i=1}^n x_i = 0$.

II. Proposed Method

This section describes the proposed PWFP feature selection method in detail. Before going to the details of the proposed method, a brief discussion of the existing methods is provided which stand as the ground for our arguments.

The feature selection problem may be formulated as finding m features out of d dimensions, which will provide the optimum classification accuracy. Thus, it involves $\binom{d}{m}$ candidates and becomes a combinational optimization problem, the solution to which is very challenging. Usually, heuristic strategy [6] is used to find the best features by evaluating each feature independently. In this case, the evaluation is carried for each d features (thus, d candidates), and the top m features are selected. Most of the available feature selection methods in the literature try to impose different evaluation criteria to obtain suitable performance.

Two widely used filter-based feature selection methods are FS and LS. The evaluation criterion used in FS [6] maximizes the between-class variance while minimizing the within-class variance. Thus, the FS is formulated as

$$F(f^i) = \frac{\sum_{k=1}^c n_k(\mu_k^i - \mu^i)^2}{\sum_{k=1}^c n_k(\sigma_k^i)^2} \tag{1}$$

where μ_i is the overall mean of ith feature, μ_k^i is the ith feature mean of kth class, σ_k^i is the ith feature variance of kth class, and n_k is the number of samples of kth class. On the other hand, LS [9] takes into account the similarity or the closeness of the data points for feature evaluation. It constructs a graph to reflect the local geometric structure and seeks the features which respect that graph. It tries to find the feature that minimizes the difference of the data points from the same class or the closely situated points, while possessing a high global variance. Given the similarity score between points x_j and x_k as S_{jk}, the LS is calculated by,

$$L(f^i) = \frac{\sum_{j,k}(f^{ij} - f^{ik})^2 S_{jk}}{Var(f^i)} \tag{2}$$

where $Var(.)$ represents the variance of the feature. Usually, Euclidean distance is used to find the similarity for graph construction. However, the computation of Euclidean distance has its inherent problems. When the dimension of the data points is very high, the distance metrics become meaningless [16]. Thus, finding close points in $\mathbb{R}^d$ is a difficult task and it affects the similarity measures based on the Euclidean distance and the local graph structure as well. The Euclidean distance is given by,

$$dist^2(x_j, x_k) = \sum_{i=1}^d (x_j^i - x_k^i)^2 = (x_j - x_k)^{\mathrm{T}}(x_j - x_k) \tag{3}$$

which uses the square of the difference of each feature dimension to compute the distance. Consider two points closely situated in high dimensional space. The presence of noise in any one dimension will increase the distance between these pair of points. Even normalization of features does not help much to alleviate the issue. And the computation of FS, LS, and other such methods are also affected.

When the dimensionality of data points is very high compared to the number of samples available for training, these methods experience a few disadvantages. First, the high dimensional data are almost empty and the computed variance carries no meaning unless the samples represent the data distribution properly. Second, the distance measure for these high dimensional data becomes almost uniform. Therefore, the similarity based methods are adversely affected. Third, the graph-structures formed with an insufficient number of instances are inaccurate to represent the feature manifold. Finally, the features selected by these heuristic algorithms are sub-optimal as each feature is computed independently and

ISBN 978-0-9928626-7-1 © EURASIP 2017

the effects of the combination of more than one feature are neglected. For example, two features may have low individual scores, however, their combined score may be very high. In this case, FS will not select either of them, although they should be selected for accurate classification.

Our formulation is based on the idea of Fisher score computation. We propose a naive way of selecting features based on pair-wise feature similarity. Feature variance should be low for the points belonging to the same class, while it should be high for points belonging to different classes. Fisher criterion uses similar logic, while considering all the samples of different classes altogether. However, the computation of mean and variance are affected with low sample size. Therefore, we use the pair-wise feature similarity to select the appropriate features.

A feature is said to be a 'good' feature if it keeps the samples of the same class close, while keeping the points from different classes far away. Alternatively, if we consider a pair of points, a good feature should have the following properties. 1) For the pair that belongs to the same class, the values of feature should be close. 2) For the pairs belonging to different classes, the feature should be able to differentiate the classes easily. Thus, we seek the feature dimensions along which the point pairs are very close for the same class and very far for different classes.

Lets define $p_{jk} = [b_1, b_2, ..., b_d]^{\mathrm{T}}, b_i \in \{0, 1\}$, with $b_i = 1$ as the features along which a pair of points $(x_j, x_k); y_j = y_k$ are close to each other. Thus, the $b_i = 1$ features in p_{jk} are the features along which the pair-wise with-in variance is minimum. We can choose these features by sorting the distance between individual features in ascending order and selecting the first few features. One naive way of doing so is to choose the features satisfying the following optimization problem,

$$\begin{aligned} &\max_{p_{jk}} \; p_{jk}^T p_{jk} \\ &s.t. \; |x_j - x_k|^{\mathrm{T}} p_{jk} < \tau \end{aligned} \tag{4}$$

where τ is a threshold. Here we use manhattan distance metric as it has been reported to have significant performance [17] for high dimensional data. Suppose, we need to keep β number of features out of d, which are close for the pair (x_j, x_k). This even makes the selection process more easy.

$$\min_{p_{jk}} \; |x_j - x_k|^{\mathrm{T}} p_{jk} \; ; \quad s.t. \quad p_{jk}^T p_{jk} = \beta \tag{5}$$

The manhattan distance between x_j and x_k is $\sum_{i=1}^{d} |x_j^i - x_k^i| = |x_j - x_k|^{\mathrm{T}} \mathbf{1}$, where $\mathbf{1}$ is a vector of ones. Thus, we can interpret the term $|x_j - x_k|^{\mathrm{T}} p_{jk}$ as a distance measure with a sub-set of features for which $b = 1$.

Similarly, let $q_{jk} = [b_1, b_2, ..., b_d]^{\mathrm{T}}, b_i \in \{0, 1\}$ be the features along which the pair $(x_j, x_k); y_j \neq y_k$ are farthest if $b_i = 1$. A similar way of finding the features that discriminate the points from different classes can be given by,

$$\max_{q_{jk}} \; |x_j - x_k|^{\mathrm{T}} q_{jk} \; ; \quad s.t \quad q_{jk}^T q_{jk} = \beta \tag{6}$$

We collect the information from all such possible pairs which are represented by P and Q respectively, given by

$$\begin{aligned} P &= \frac{1}{N_p} \sum_{j,k; y_j = y_k} p_{jk} \\ Q &= \frac{1}{N_q} \sum_{j,k; y_j \neq y_k} q_{jk} \end{aligned} \tag{7}$$

where N_p and N_q are normalization factors. We use $N_p = \sum_{k=1}^{c} \binom{n_k}{2}$ and $N_q = \sum_{j,k; j \neq k} n_j n_k$. Here P and Q represents the normalized histogram of features based on their contribution toward closeness or discriminating power between different classes.

Now, we seek the feature dimensions which are present in both P and Q. To be specific, a good feature should be capable of discriminating between points from different classes while bringing the points of same class closer. Therefore, a good feature is a feature that has higher probability occurrence in both $P = [p^1, p^2, ..., p^d]$ and $Q = [q^1, q^2, ..., q^d]$. A reasonable criterion for choosing a good feature is to minimize the object function given by,

$$\min_i \left| \frac{p^i - q^i}{p^i + q^i} \right| \tag{8}$$

Apparently, once P and Q are found, the problem (8) can be solved by element-wise operation of these two vectors. Furthermore, the term $S(i) = \left| \frac{p^i - q^i}{p^i + q^i} \right|$ can be used as a score for selecting features. The top m features are those with the lowest scores. The algorithm for the proposed feature selection method is provided in algorithm 1.

As can be observed, we first select the optimal set of features for each pair of sample instances. These set of features from all such pairs are further brought together and the best features are selected based on the scores assigned to each feature. Thus, the proposed PWFP involves the combinational

Algorithm 1 Algorithm for feature selection based on the pair-wise feature proximity (PWFP)

Input: Training data $\{(x_i, y_i)\}_{i=1}^{n}$, the selected feature number m, and the parameter β

Output: The selected feature subset

```
for ∀(x_j, x_k) ∈ X do
    if y_j = y_k then
        Compute p_jk using (5)
    else
        Compute q_jk using (6)
    end if
end for
Compute P and Q using (7).
Calculate the score of each feature based on S(i) = |(p^i − q^i)/(p^i + q^i)|
Rank the features according to the scores in descending order
Select the leading m features
```

ISBN 978-0-9928626-7-1 © EURASIP 2017

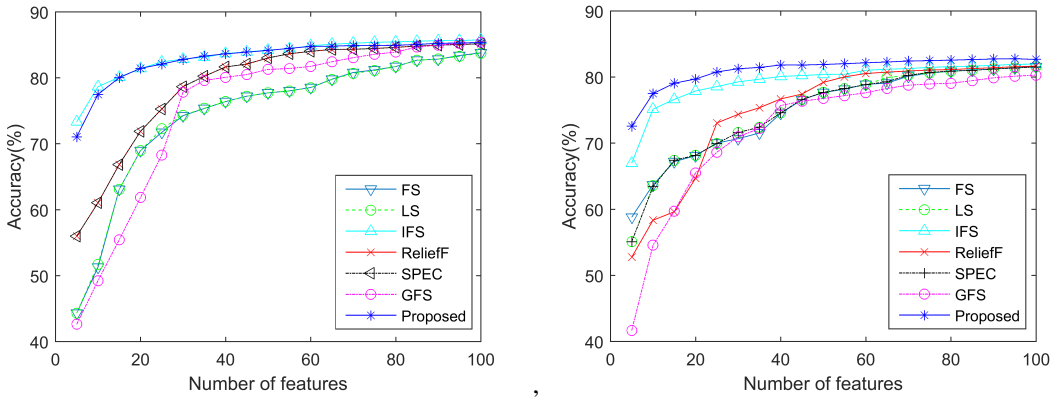

Fig. 1. Comparison of performances of proposed (PWFP) and other methods in (a) Botswana (left) and (b) Salinas (right) database. [best viewed in color]

feature selection followed by a heuristic approach of score assignment to each feature. It takes the advantages of selecting a group of features based on the pair-wise proximity in feature values and the lower computational complexity of heuristic search for assigning scores and ranking the features.

The first step in PWFP is similar to that of Fisher criterion. If we consider a pair of samples, then it is equivalent to selecting m features using FS. However, FS considers all data together, while PWFP uses the pair-wise data. Unlike FS or LS, our method selects multiple features for each pair of data and combines them to select the best ones in a later stage. Therefore, the fear of suboptimal feature selection due to the evaluation of independent features is avoided. No graph representations are considered as we have assumed less number of available labeled samples. Furthermore, PWFP does not use similarity based on full feature set, which avoids over-fitting.

III. Experiments and Results

The performance of the PWFP is validated through the experiments conducted on several real-world data sets. A preprocessing step was carried out to normalize the samples to zero mean and unit variance. In all cases, a few samples were selected for training, while the rest were used for testing purpose. Linear support vector machines were used for classification purpose in all cases. We set the value of β to be 10% of the dimension in all the experiments.

A. Hyperspectral data sets

Hyperspectral images consist of hundreds of spectral bands to provide information about the properties of land cover. With a few annotations from the experts, segmentation and classification algorithms find the labels for the rest of the image. This is a special case of high dimensional low sample size data. In our experiments, we used Botswana[1] ($d = 145, c = 14$) and Salinas scene[1] ($d = 204, c = 16$). The training set was constructed with random selection of 10 sample points from each class and the rest were treated as test data. We report the average performance over ten iterations.

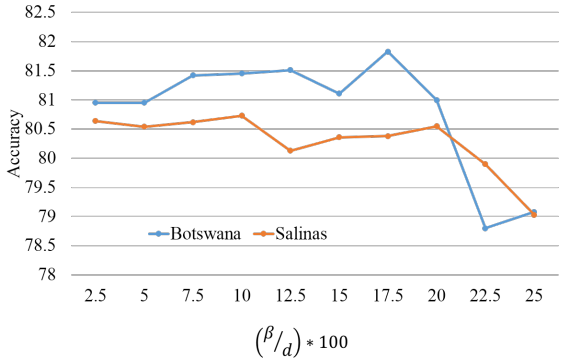

Fig. 2. Effect of variation of β on classification performance.

The recognition accuracy with respect to the number of selected features of all the feature selection methods is provided in Fig. 1. We observed that some of the existing methods performed almost equal on these data sets. The low sample size might be the reason behind different methods having similar properties. IFS performed the best among them. However, the PWFP achieved remarkable accuracy in Salinas database with a different number of selected features, while performing close to IFS in Botswana data set. The accuracy obtained with varying β is illustrated in Fig. 2. As can be observed, the accuracy increases with increase of β, peaks, and then decreases. Empirically, we selected β as 10% of the feature dimension in all the experiments.

B. Face Recognition

We carried out experiments on the ORL face recognition data set[2], which contains 10 images for each of the 40 participants ($c = 40$). The face images were resized to 32×32 and vectorized ($d = 1024$). Randomly 5 images from each person were selected for training and the rest for testing. The

[1] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

ISBN 978-0-9928626-7-1 © EURASIP 2017

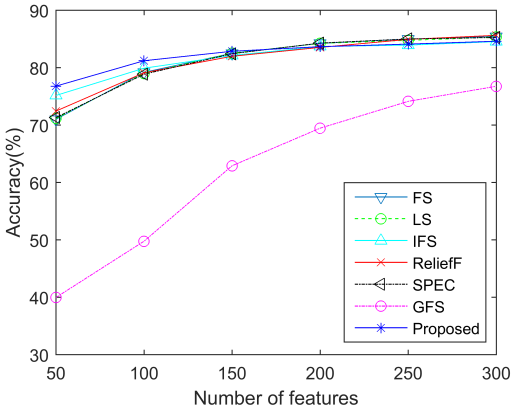

Fig. 3. Performance of different feature selection methods on ORL data set.

average of 10 experiments with a different number of feature selection is provided in Fig. 3.

As can be seen, the PWFP outperforms the other feature selection methods when a few features are selected. GFS underperformed in this data set, while the other methods achieved almost equivalent accuracy. Lack of sufficient data for each class might be the reason for the failure of GFS algorithm.

C. Other data sets

A few publicly available databases having less number of sample instances with high dimension were used in our experiments. The chosen data sets are Colon cancer diagnosis dataset[3], Lung cancer[4], protein[5] , ionosphere[5], arcene[5], and tox-171[6]. For each data set, 10% samples were randomly selected as training data and the rest were treated as test data. We repeated this procedure five times and the average performance is reported in Table I.

As can be seen in Table I, the data sets are arranged in increasing order of data dimensionality and the performance of the proposed method is compared with a few other methods that optimize the features globally. The accuracy achieved by PWFP is less when the data dimension is low (for protein and ionosphere). However, its performance goes higher as the dimensionality increases. PWFP achieved the best classification accuracy for the last four data sets. This validates the efficiency of the PWFP for high dimensional data with low sample size.

IV. Conclusion

In this paper, we propose a feature selection method based on pair-wise feature proximity. We use the closeness (or remoteness) of a feature dimension among a pair of points from the same (or different) class to select the efficient features for the class discrimination. The proposed method is analyzed extensively with a few high dimensional low sample size databases. It is found that the proposed method outperforms the existing algorithms when the database has a few samples with very high dimension.

[3]http://www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/chm/Alon.html

[4]http://www.pnas.org/content/98/24/13790/suppl/DC1

[5]http://archive.ics.uci.edu/ml/datasets.html

[6]http://featureselection.asu.edu/old/datasets.php

TABLE I
Classification results on different data sets when 10% data are used for training and the number of selected features is set to be 50% of the dimensionality of the data.

Methods	**protein**	**ionosphere**	**colon**	**lung**	**TOX-171**	**arcene***
(samples, dimensions)	(116,20)	(351,34)	(62,2000)	(203, 3312)	(171, 5748)	(200, 10000)
IFS	36.34	79.74	58.9	83.4	55.94	75.8
SPEC	**45.57**	81.71	58.54	84.17	56.33	56
GFS	43.65	**81.96**	60.72	75.05	57.12	79.6
Proposed	43.26	78.96	**64.36**	**84.72**	**57.77**	**81**

* For arcene data set, 50% data were used for training.

References

[1] M. S. Cheema, A. Eweiwi, and C. Bauckhage, "High dimensional low sample size activity recognition using geometric classifiers," *Digital Signal Processing*, vol. 42, pp. 61–69, 2015.

[2] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classification: Algorithms and Applications*, p. 37, 2014.

[3] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[4] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 3, pp. 619–632, 2013.

[5] M. Zaffalon and M. Hutter, "Robust feature selection using distributions of mutual information," in *18th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2002, pp. 577–584.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2001.

[7] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines." in *International Conference on Machine Learning (ICML)*, vol. 98, 1998, pp. 82–90.

[8] H. Liu and H. Motoda, *Computational methods of feature selection*. CRC Press, 2007.

[9] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *NIPS*, vol. 186, 2005, p. 189.

[10] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection." in *AAAI Conference on Artificial Intelligence*, vol. 2, 2008, pp. 671–676.

[11] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *24th International Conference on Machine Learning (ICML)*. ACM, 2007, pp. 1151–1157.

[12] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4202–4210.

[13] C. Liu, W. Wang, Q. Zhao, X. Shen, and M. Konan, "A new feature selection method based on a validity index of feature subset," *Pattern Recognition Letters*, vol. 92, pp. 1–8, 2017.

[14] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *27th International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 266–273, 2011.

[15] Y. Luo, D. Tao, C. Xu, D. Li, and C. Xu, "Vector-valued multi-view semi-supervsed learning for multi-label image classification," in *AAAI Conference on Artificial Intelligence*, 2013, pp. 647–653.

[16] C.-M. Hsu and M.-S. Chen, "On the design and applicability of distance functions in high-dimensional data space," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 4, pp. 523–536, 2009.

[17] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International Conference on Database Theory*. Springer, 2001, pp. 420–434.

ISBN 978-0-9928626-7-1 © EURASIP 2017