Clustering Multivariate Empiric Characteristic Functions for Multi-Class SVM Classification

María-Dolores Cubiles-de-la-Vega, Rafael Pino-Mejías, and Esther-Lydia Silva-Ramírez

***Abstract*—**A dissimilarity measure between the empiric characteristic functions of the subsamples associated to the different classes in a multivariate data set is proposed. This measure can be efficiently computed, and it depends on all the cases of each class. It may be used to find groups of similar classes, which could be joined for further analysis, or it could be employed to perform an agglomerative hierarchical cluster analysis of the set of classes. The final tree can serve to build a family of binary classification models, offering an alternative approach to the multi-class SVM problem. We have tested this dendrogram based SVM approach with the one-against-one SVM approach over four publicly available data sets, three of them being microarray data. Both performances have been found equivalent, but the first solution requires a smaller number of binary SVM models.

***Keywords*—**Cluster Analysis, Empiric Characteristic Function, Multi-class SVM, R.

I. Introduction

SUPPORT VECTOR MACHINES (SVM) are a powerful family of supervised machine learning techniques. They emerged from Statistical Learning Theory, or Vapnik-Chernovenkis theory [1]-[3] and several extensions were successively proposed. When used for a two-class classification problem where the set of binary labeled training patterns is linearly separable, the SVM separates both classes with a hyper-plane that is maximally distant from them (“the maximal margin hyper-plane”). If linear separation is not possible, the feature space is enlarged using basis expansions such as polynomials or splines. Moreover, explicit specification of this transformation is not necessary, as a kernel function that computes inner products in the transformed space can be employed**.**

SVM and its variants have been successfully applied in many domains, for example in two-class classification of microarray data [4]-[6]. Bioinformatics data sets usually contain measurements for thousands of genes, which proves problematic for many traditional methods, while SVM are well suited to obtain classification models with such high dimensional data.

The extension of the binary SVM model to the multi-class scenario is still a research topic. In general it is computationally more expensive to solve a multi-class problem than a binary model with the same number of data. A usual approach is based on the construction of a set of binary SVM models. For example, the one-against-all method, (for example [7]) builds *M* binary SVM models for a problem with *M* classes, where the *i*-th model tries to separate the class i from the remaining categories. Thus, the classification rule for each model is based on the sign of a decision function $m_i(x)$. The final decision is based on the class which has the largest value of the decision functions $m_1(x),\ldots,m_M(x)$.

Another approach is the one-against-one method, initially introduced in [8] for neural networks, where $M(M-1)/2$ models are obtained, one for each pair of classes, and a voting scheme provides the final decision.

The directed acyclic graph SVM (DAGSVM) proposed in [9] uses a rooted binary directed acyclic graph which has internal nodes and leaves. Each node is a binary SVM of i-th and j-th classes. As it is explained in [10], given an input vector to be classified, starting at the root node, the binary decision function is evaluated. Then it moves to either left or right depending on the output value, finally reaching a leaf node which indicates the predicted class.

The Dendrogram-based SVM model (DSVM) [11] is an alternative based on the previous realization of a hierarchical cluster analysis of the *M* classes. In each level of the dendrogram, a binary classification problem is formulated to separate two groups of classes. The final decision is computed by presenting the input vector to the set of *M*-1 SVM models in a tree decision form, until an assignation to a class is reached. However, the distance between classes is computed in [11] as the distance between the *M* gravity centers. It is well known that the arithmetic mean can be a bad representative of a distribution, for example when there exist outliers in the sample, or when asymmetric distributions are obtained. We propose a dissimilarity measure between the multivariate empiric characteristic functions of the *M* samples. It can be used to identify and group similar classes, or either it can serve to perform the cluster analysis of the classes. This measure is presented in section III. Previously, section II presents the SVM model as it is available in the R system. Section IV contains the application of this multi-class SVM approach to four public data sets. Finally, section V contains the main conclusions and future work.

María-Dolores Cubiles-de-la Vega is with the Department of Statistics and Operational Research, University of Seville, Spain, (e-mail: cubiles@us.es).

Rafael Pino-Mejías is with the Department of Statistics and Operational Research, University of Seville, Spain, (e-mail: rafaelp@us.es).

Esther-Lydia Silva-Ramírez is with the Department of Language and Computer Sciences, University of Cadiz, Spain, (e-mail: esther.silva@uca.es).

This work was partially funded by Institute of Statistics of Andalusia, grants 200900589 and 201000887.

II. TWO-CLASS SVM MODEL

We have fitted the SVM models in section IV with the *svm* function available in the library e1071 [12] of the R system [13], which offers an interface to the award-winning C++ implementation, LIBSVM, by Chan and Lin. For a binary problem, the data set is described by n training vectors $\{x_i, y_i\}$, $i=1,2,\ldots,n$, where the p-dimensional vectors x_i contain the predictor features and the n labels $y_i \in \{-1,1\}$ identify the class of each vector. Among the several variants of SVM existing in the library e1071, we have used C-classification with the Radial Basis Kernel. The primary quadratic programming problem to be solved is:

$$\begin{aligned} &\underset{\mathbf{w},b,\xi}{Min} \quad \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ &y_i(\mathbf{w}^t\phi(\mathbf{x}_i)+b) \geq 1-\xi_i \\ &\xi_i \geq 0,\ i=1,2,\ldots,n \end{aligned} \tag{1}$$

$C>0$ is a parameter controlling the trade-off between margin and error. The dual problem is

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{Min} \quad \frac{1}{2}\boldsymbol{\alpha}^t R\boldsymbol{\alpha} - \mathbf{e}^t\boldsymbol{\alpha} \\ &0 \leq \alpha_i \leq C, \quad i=1,2,\ldots,n \\ &\mathbf{y}^t\boldsymbol{\alpha} = 0 \end{aligned} \tag{2}$$

where e is the n-vector of all ones, and R is a positive semi definite matrix defined by $R_{ij}=y_iy_jK(x_i,x_j)$, $i,j=1,2,\ldots,n$, being $K(x_i,x_j)=\phi(x_i)\phi(x_j)$ the kernel function. A vector x is classified by the decision function

$$sign\left(\sum_{i=1}^{n} y_i\alpha_i K(\mathbf{x}_i, x) + b,\right) \tag{3}$$

depending on the margins

$$m_i = \sum_{i=1}^{n} y_i\alpha_i K(\mathbf{x}_i, x) + b, \quad i=1,2,\ldots,n \tag{4}$$

The Radial Basis Function (RBF) was our choice for K:

$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\gamma\|\mathbf{u}-\mathbf{v}\|^2\right) \tag{5}$$

So, two parameters was tuned in the experiments described in section IV: C and γ. We adopted the suggestions of the authors of LIBSVM [14] about the definition of the search grid. Thus, our search for C included small and big values, while the explored values for γ have been selected around the default value of the *svm* function in the R library e1071, defined as $1/p$, being p the number of predictors. This search was performed through a cross validation procedure with the *tune.svm* function also available in the library e1071.

III. A DISSIMILARITY MEASURE BETWEEN EMPIRIC CHARACTERISTIC FUNCTIONS

Let $X_1,\ldots,X_n$ be a multivariate sample from a p-dimensional continuous population, and let F_n be the empirical cumulative distribution function, defined by $F_n(x)=N(x)/n$, where $N(x)$ is the number of $X_j \leq x$. Following [15] the empiric characteristic function φ is defined for any p-dimensional real vector t as:

$$\varphi(t) = \int_{x\in R^p} e^{it'x}dF_n(x) = \frac{1}{n}\sum_{j=1}^{n} e^{it'X_j} \tag{6}$$

where t' denotes the transpose of the column vector t. The empiric characteristic function has several important statistical properties [15]: it allows an easy characterization of independence and symmetry, it retains all the information existing in the sample and it can be efficiently computed. Thus, several inferential procedures based on this function have been proposed.

Now we consider the M samples from the M populations appearing in the multi-class problem. Let $G_1,\ldots,G_M$ be the corresponding multivariate empirical distribution functions. Thus, the M multivariate empirical characteristic functions are defined as follows, for $j=1,2,..,M$, where $X_{j,r}$ is the p-sized column vector corresponding to the r element of the n_j sized sample j:

$$\phi_j(t) = \int_{x\in R^p} e^{it'x}dG_j(x) = \frac{1}{n_j}\sum_{r=1}^{n_j} e^{it'X_{j,r}} \tag{7}$$

Fixed t, let $d_{ij}(t)^{1/2}$ the Euclidean distance between the row complex vectors associated to the values that i-th and j-th empiric functions take in t:

$$d_{ij}(t)^{1/2} = \|\phi_i(t)-\phi_j(t)\| = \sqrt{(\phi_i(t)-\phi_j(t))'(\phi_i(t)-\phi_j(t))} \tag{8}$$

Then, it is easy to verify that

$$\begin{aligned} d_{ij}(t) &= \|\phi_i(t)-\phi_j(t)\|^2 = \frac{1}{n_i}+\frac{1}{n_j}+ \\ &+\frac{1}{n_i^2}\sum_{\substack{r,s=1\\ r\neq s}}^{n_i}\cos(t'(X_{i,r}-X_{i,s}))+ \\ &+\frac{1}{n_j^2}\sum_{\substack{r,s=1\\ r\neq s}}^{n_j}\cos(t'(X_{j,r}-X_{j,s}))+ \\ &-\frac{2}{n_in_j}\sum_{r=1}^{n_i}\sum_{s=1}^{n_j}\cos(t'(X_{i,r}-X_{j,s})) \end{aligned} \tag{9}$$

We now consider the orthonormal basis *B* of the *p*-euclidean space:

$$B=\{t_\eta=(0,0,...^{\eta-1},0,1,0,...0')',\eta=1,2,...,p\} \tag{10}$$

We define the following measure of dissimilarity between the *i*-th and *j*-th empiric characteristic functions, based on *B*, and therefore, between their corresponding classes:

$$D(i,j)=\sum_{\eta=1}^{p} d_{ij}(t_\eta) \tag{11}$$

Thus, a dissimilarity matrix *D* can be computed to measure the distances between the *M* classes, helping to clarify the existing relations between the different classes. For example, the two most dissimilar (or most similar) classes can be detected. Another possibility is to perform a cluster analysis to obtain a taxonomy of the *M* classes, and to build a set of *M*-1 binary classification models. For a hierarchical cluster analysis, in each step the dissimilarities matrix must be recomputed. This can be easily implemented using the following tips.

First,

$$d_{ij}(t_\eta)=(T_{i,\eta}-T_{j,\eta})^2+(S_{i,\eta}-S_{j,\eta})^2 \tag{12}$$

Being

$$T_{i,\eta}=\frac{1}{n_i}\sum_{r=1}^{n_i}\cos(X_{i,r}^{\eta}) \quad ST_{i,\eta}=\frac{1}{n_i}\sum_{r=1}^{n_i}\sin(X_{i,r}^{\eta}) \tag{13}$$

So we have

$$D(i,j)=(T_{(i)}-T_{(j)})(T_{(i)}-T_{(j)})'+(S_{(i)}-S_{(j)})(S_{(i)}-S_{(j)})' \tag{14}$$

with

$$T_{(i)}=(T_{i,1},....,T_{i,p})' \quad S_{(i)}=(S_{i,1},....,S_{i,p})' \tag{15}$$

Last vectors can be loaded in two matrices *T* and *S*:

$$T=(T_{(1)}^{'},....,T_{(p)}^{'}) \quad S=(S_{(1)}^{'},....,S_{(p)}^{'}) \tag{16}$$

When clusters *i* and *j* are joined, *T* and *S* are immediately updated by

$$\begin{aligned} T_{(ij)}&=\frac{1}{n_i+n_j}(n_iT_{(i)}+n_jT_{(j)}) \\ S_{(ij)}&=\frac{1}{n_i+n_j}(n_iS_{(i)}+n_jS_{(j)}) \end{aligned} \tag{17}$$

TABLE I
DISTANCES BETWEEN THE EMPIRIC CHARACTERISTIC FUNCTIONS FOR THE KHAN DATA SET

	BL	EW	NB
EW	207.4		
NB	209.5	133.1	
RM	229.8	95.4	119.0

This way, the dissimilarity matrix can be efficiently actualized during the clustering process.

IV. NUMERICAL EXPERIMENTS

We have compared the multi-class dendrogram SVM model based on the previously defined dissimilarity matrix, which we will call DSVMC, and the one-against-one method. For each fitted SVM model,. a grid search for an appropriate configuration of *C* and *γ* (parameter of the radial basis kernel in the *svm* function) was realized by 10-fold cross-validation with the aid of the *tune.svm* function in the e1071 library in R.

A. Khan data set

This data set about the small, round blue cell tumors (SRBCTs) of childhood includes 63 samples classified as neuroblastoma (NB), rhabdomyosarcoma (RM), Burkitt lymphomas (BL) and the Ewing family of tumors (EW). 25 test samples are also available. Data from the cDNA microarray experiment contains 2308 genes. [16] used this data set, and it can be downloaded at the URL in [17], although we used the accompanying disk in [18]. The data was log transformed and normalized, and the 200 top genes according to *p*-values corresponding to the one-way ANOVA were selected. Table I contains the dissimilarity matrix between the four classes, based on the empiric characteristic functions, computed on the training data set.

Fig. 1 displays the clustering process, and the three associated binary SVM models, denoted by M1, M2 and M3. Thus, given a case to be classified, its corresponding 200 gene expression values are the input to M1. If the decision of M1 is BL, that is the classification, otherwise, the case is entered into M2. If M2 says NB, this is the decision, otherwise M3 outputs a final classification.

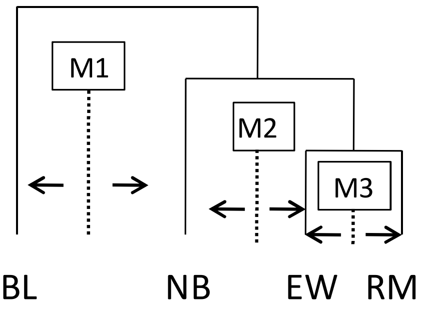

Fig. 1 Cluster analysis of the classes of Khan data set

The one-against-one SVM model was fitted on the 63 training samples, and the estimated error rate on the test set was 0%. The DSVMC model also provided a zero test error rate. Therefore, both models provided equivalent and good performances, although DSVMC only needs three binary SVM models, while the usual one-against-one SVM approach requires six binary SVM models.

B. DNA data set

This primate splice-junction gene sequences data set consists of 3186 data points and 180 indicator binary variables. The problem is to recognize the 3 classes (ei, ie, neither), i.e., the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). We have used the available version in the *mlbench* library in R [19]. We must note that this data set has a dimensionality different to the typical microarray data sets, and therefore it can illustrate the performance of DSVMC for those situations. 50 random splits into training (2000 samples) and test sets (1186 samples) were performed.

The 50 dendrograms were all equal, as it is presented in fig. 2. M1 tries to separate ie class from ei and n clasess, while M2 is designed to discriminate between ei and n. From the 50 resulting M1 and M2 models, the obtained mean test error rate was 4.18% for DSVMC, while the one-against-one SVM model provided a slightly greater value of 4.36%. A paired two sample t-test for the null hypothesis of population means equality was realized, detecting a weak evidence in favor of DSVMC (p-value = 0.07197).

Fig. 3 displays a box and whisker plot of the 50 differences between the test error rates for the SVM one-against-one model and the DSVMC method. 65% of the random splits provided a test DSVMC error rate lower than SVM one-against-one.

C. Nci data set

This data set comprises the expression matrix of 6830 genes and 64 samples, for patients suffering a tumor. Eight types of tumor are present: NSCLC, OVARIAN, CNS, RENAL, COLON, LEUKEMIA, BREAST and MELANOMA. We used the *nci* data set available in the *ElemStatLearn* library in R [20]. A similar preprocessing as in Khan data set was performed, selecting the 200 top genes. 100 random splits into training (48 samples) and test sets (16 samples) were also performed.

The 100 obtained dendrograms were very different, unlike DNA data set. The mean test error rate for DSVMC was 21.66%, while one-against-one SVM provided 21.52%. A paired two sample *t*-test for the null hypothesis of population means equality was realized, accepting the equality (p-value = 0.51)

TABLE II
MEAN VALUES OF DISTANCES BETWEEN THE EMPIRIC CHARACTERISTIC FUNCTIONS FOR THE VEHICLE DATA SET

	BUS	OPEL	SAAB
OPEL	1.17		
SAAB	1.22	0.22	
VAN	1.17	0.32	0.29

The reduction in the number of binary models which DSVMC offers for this data set is very important, only 7 models, while one-against-one needs 28 binary SVM models.

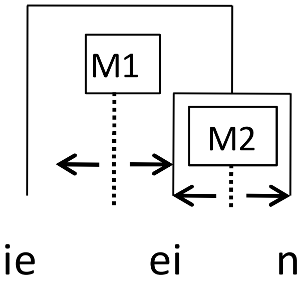

Fig. 2 Cluster analysis of the classes of DNA data set

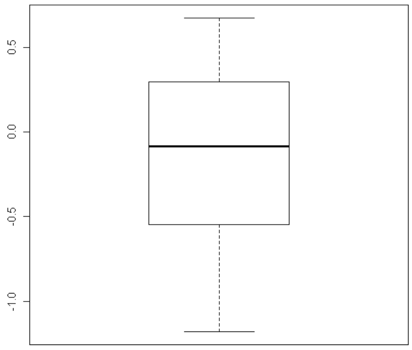

Fig. 3 Differences between the 50 test error rates for the DSVMC and one-against one SVM. DNA data set

D. Vehicle data set

This data set is available in the library *mlbench* of R [19]. The purpose is to classify a given silhouette as one of four types of vehicle, using a set of 18 features extracted from the silhouette.

The 846 cases were randomly split into training (75%) and test (25%) sets. This random split was independently repeated 100 times. Table II contains the mean values of D over the 100 iterations.

Fig. 4 shows the three binary SVM models resulting from the hierarquical agglomerative clustering process in 95 pf the 100 splits. M1 is fitted to separate bus from the aggregated class {van, opel, saab}. M2 is designed to discriminate between van and {opel, saab}. M3 discriminates between opel and saab.

The mean test error rate for DSVMC was 15.64%, while one-against-one SVM provided 16.03%. A paired two sample *t*-test for the null hypothesis of population means equality was again realized, accepting the equality (p-value = 0.52).

Again, the equivalent performance is accompanied by a lower number of models both in training and testing phases when using DSVMC, which at most requires three binary models, while one-against-one classification always needs the six binary models.

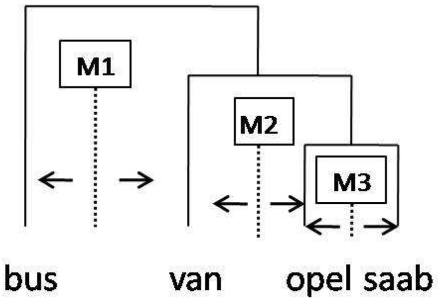

Fig. 4 Cluster analysis of the classes of the Vehicle data set

V. CONCLUSION

A dissimilarity measure between the subsamples appearing in a multivariate dataset has been presented. It is based on the multivariate empiric characteristic functions, all the information existing in the sample is used, and it can be efficiently computed. This measure can be used to obtain a clustering description of the different subsamples. We have tested its use to derive a set of binary SVM models as an alternative inside the multi-class problem. The empirical results suggest a similar performance to the SVM one-against-one but it only requires *M*-1 binary models for a problem with M classes. Future works could include other multi-class SVM approaches, the use of categorical variables and a wider empirical study. Other models, as logistic regression, could be fitted to obtain the set of binary classification models, defining an alternative approach to multinomial logistic regression.

REFERENCES

[1] B.E. Boser, I.M. Guyon, and V.N. Vapnik, “A training algorithm for optimal margin classifiers”, in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, ACM Press, Pittsburgh, 1992, pp. 144-152.

[2] V. Vapnik, Statistical Learning Theory, John Wiley, New York, 1998.

[3] N. Cristianini, J. Shawe-Taylor. An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2002.

[4] I. Guyon, J. Weston, S. Barnhill, V. Vapnik . Gene selection for cancer classification using support vector machines, Machine Learning, 46(1): 389-422, 2002.

[5] L. Wang, J. Zhu, H. Zou. Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics 24(3): 412-419, 2008.

[6] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani. 1-norm support vector machines. Advances in Neural Information Processing Systems 16(1): 49-56, 2004.

[7] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of classifier methods: A case study in handwriting digit recognition, in Proceedings of the International Conference on Pattern Recognition, 1994, pp. 77–87.

[8] A.S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited: A stepwise procedure for building and training a neural network, in Neurocomputing: Algorithms, Architectures and Applications, J. Fogelman, Ed. New York: Springer-Verlag, 1990.

[9] J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAG’s for multiclass classification, in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000, vol. 12, pp.547–553.

[10] C.W. Hsu, and C.J. Lin. A comparison of Methods for Multiclass Support Vector Machines, IEEE Transactions on Neural Networks, 13(2), pp.415–425, 2002.

[11] K.Benabdeslem, and Y. Bennani. Dendrogram-based SVM for Multi-Class Classification. Journal of Computing and Information Technology, 14(4) pp. 283–286, 2006.

[12] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer and and A. Weingessel. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-18, 2008.

[13] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2012.

[14] C.C., Chang, and C.J. Lin. LIBSVM: a library for support vector machines.URL: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz, 2001.

[15] A. Feuerverger, R.A. Murieka. The empiric characteristic function and its application, The Annals of Statistics 5, 88-97, 1977.

[16] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Atonescu, C. Peterson, P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Med. 7, 673–679, 2001.

[17] http:// research.nhgri.nih.gov/microarray/Supplement/Images/ supplemental_data.

[18] S. Deshmukh, S. Purohit. Microarray data. Statistical Analysis Using R, Alpha Science International Ltd., Oxford, 2007.

[19] F. Leisch, E. Dimitriadou. mlbench: Machine Learning Benchmark Problems. R package version 1.1-6, 2009.

[20] Material from the book's webpage, R port and packaging by Kjetil Halvorsen . ElemStatLearn: Data sets, functions and examples from the book: "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman. R package version 0.1-6. 2007.