Asymmetric–margin support vector machines for lung tissue classification

Jimison Iavindrasana, Adrien Depeursinge, Gilles Cohen, Antoine Geissbuhler and Henning Müller

***Abstract*— This paper concerns lung tissue classification using asymmetric–margin support vector machine (ASVM) to handle the imbalance of the positive and negative classes in a one–against–all multiclass classification problem. The hyperparameters of the algorithm are obtained using an optimization of the upper bound of the leave–one–out error of the ASVM. The ASVM is applied on the dataset with its original distribution and oversampled so that the ratio of the examples is equal to the prevalence of patients having the tissue in the database. The two versions of the ASVM models were compared with a model build with a conventional SVM. The ASVM improved the results obtained with a conventional SVM. The incorporation of prior knowledge concerning the prevalence of the patients improved the results obtained with ASVM.**

I. Introduction

Interstitial lung diseases (ILD) form a heterogeneous group of diseases containing more than 150 disorders of the lung tissue. Many of the diseases are rare and present unspecific symptoms. During the diagnostic process, all available information including the patient's personal data, medication, past medical history, host risk factors and laboratory tests (e.g. pulmonary function tests, hematocrit, ...) are meticulously analyzed to find any indicator of the presence of an ILD. Beside the patient's clinical data, imaging of the chest allows to resolve an ambiguity in a large number of cases by enabling the visual assessment of the lung tissue [1]. The most common imaging modality used is the chest X–ray because of its low cost and radiation dose. It is of sometimes of limited usefulness for the characterization of lung tissue as these are overlaid with other anatomical structures, making the reading sometimes difficult. The gold standard imaging technique used in case of doubt is the high–resolution computed tomography (HRCT), which provides three–dimensional images of the lung tissue with high spatial resolution. Most of the histological diagnoses of ILDs are associated with a given combination of image findings (i.e. abnormal lung tissue) [2]. The most common lung tissue patterns are *emphysema*, *ground glass*, *fibrosis*, *micronodules* and *consolidation*. These are characterized by distinct texture properties in HRCT imaging. The detection and characterization of the lung tissue patterns in HRCT is time–consuming and requires experience. In order to reduce the risk of omission of important tissue lesions and to ensure the reproducibility of image interpretation, computer–aided diagnosis (CAD) was proposed several times for HRCT of the lung [3], [4], [5], [6], [7], [8], [9], [10]. The typical approaches use supervised machine learning to draw decision boundaries in feature spaces spanned by texture attributes. The reported performance of these approaches suggests that these systems have the potential to be valuable tools in clinical routine by providing second opinions to the clinicians. However, the CAD system must include a sufficient number of classes of lung tissue to cover the heterogeneous visual findings associated with ILDs. A CAD system, which aims at detecting one single lung tissue pattern is of limited use as the radiologist still needs to look for other pathological lung tissue patterns in the image series.

The authors are with the Service of Medical Informatics, University Hospitals of Geneva, Rue Gabrielle–Perret–Gentil 4, 1211 Geneva 14, Switzerland (phone: +41 22 372 8874; email: jimison.iavindrasana@sim.hcuge.ch). Henning Müller is also with the HES–SO Valais, TechnoArk 3, 3960 Sierre, Switzerland

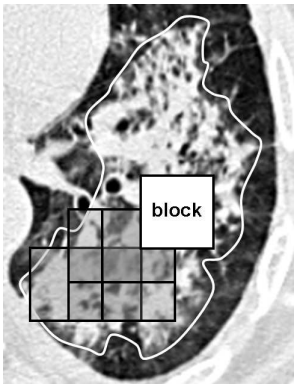

Fig. 1. Construction of the block instances from manually delineated ROIs.

A major performance problem of multi–class CAD systems is the challenge of learning from highly imbalanced datasets. In a given dataset of HRCT images of patients affected with ILDs, the instances consist of manually delineated regions of interest (ROI) showing examples of the lung tissue patterns that are cut out into square blocks that may overlap or not (see Figure 1). The resulting distributions of the classes are depending both on the prevalence of each lung tissue sort and the average ROI surface and can as a consequence be highly imbalanced. Although equal sensitivity and specificity are needed among the classes, most of the machine learning techniques favor performance of the majority class and research efforts are needed to ensure balanced performances among all classes.

In this article, support vector machines with asymmetric margins (ASVM) are used to classify the ROIs.

The paper is structured as follows: Section 2 introduces the method for handling imbalanced datasets with SVMs and the estimation of the SVM/ASVM hyperparameters. Section

3 details the materials and the classification method used followed by the presentation of the results in Section 4. A discussion of the results is found in Section 5 and a conclusion and future ideas in Section 6.

II. Learning imbalanced data sets with ASVM

After a short review of the techniques to handle imbalanced datasets the SVM algorithm is introduced followed by details on the built–in dataset imbalance management and a method to estimate the SVM/ASVM hyperparameters.

A. Imbalanced learning approaches in the literature

Two main approaches were proposed in the literature to manage imbalanced datasets: the resampling strategy and the algorithm–based strategy [11]. The first one is a data–driven strategy and performed by down–sampling the majority class or oversampling the minority class. This method has many variants with respect to the resampling technique: resampling at random, undersampling by removing redundant or noisy majority examples [12], oversampling with synthetic examples drawn using clustering algorithms [13] and [14], oversampling positive examples located near the decision function [15]. A comparative study of the available resampling strategies was carried out by [16] with the C4.5 algorithm and the random undersampling and oversampling methods outperforming all resampling strategies. The second strategy (algorithm–driven) consists of altering the misclassification cost of the classes such as in [17] or altering the data representation to achieve a high separability of the data (see for example [18]). A comparison of the resampling and cost–sensitive strategy with SVM can be found in [14]. In this comparative study, the SVM with asymmetric margins was used and it outperformed the resampling technique (combination of undersampling and oversampling with artificial examples generated with the k–means algorithm).

B. Support Vector Machines

A maximum margin classifier looks for an optimal hyperplane separating the training dataset such as the distance of the training points to the optimal hyperplane is maximized. This assumes that the training data are separable. Finding this optimal hyperplane is equivalent to resolving the following quadratic optimization problem:

$$min \qquad (w^T w) \qquad s.t. \;\; y_i(w^T x_i + b) \geq 1 \tag{1}$$

where w is a vector perpendicular to the hyperplane, b is a scalar value, $\{x_i, y_i\}_{i=1}^N$ are the training points ($x_i \in \Re^d, y_i \in \{-1, 1\}, N$ the number of examples and d the numbers of variables). If certain conditions hold and using the Lagrangian formulation, the previous problem is equivalent to its dual, which is a quadratic optimization problem and which can be solved using several optimization techniques:

$$\begin{array}{cc} max_{\alpha \geq 0} & \theta(\alpha) \;=\; \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j \\ s.t. & \sum_i \alpha_i y_i = 0 \end{array} \tag{2}$$

where α_i are the Lagrangian multipliers. A support vector machine (SVM) is a maximum margin classifier, which uses only the points on both sides of the margin or support vectors (points x_i for which $\alpha_i > 0$) to build a model.

For a non–separable training data set, penalty variables ξ_i are introduced to soften the constraints of the maximum margin formulation (1). The penalty variables ξ_i are drawn as follows: $0 < \xi_i \leq 1$ if the points are on the correct side of the hyperplane and $\xi > 1$ if the point is on the wrong side. A cost variable C is also introduced to control the trade–off between the width of the margin and the points within the margin. The final goal of the SVM classifier is then to maximize the margin while minimizing the total sum of the penalties and thus equation (1) becomes:

$$\begin{array}{cc} min & (w^T w) + C \sum_i \xi_i^p \\ s.t. & y_i(w^T x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0 \end{array} \tag{3}$$

The primal formulation of the SVM can be solved using [19] for separable cases (Eq. (1)) and [20] and for non–separable cases (Eq. (2)). However, the dual problem is often solved because the duality theory provides a convenient way to deal with the constraints. The dual optimization problem can also be written in terms of the dot products permitting the use of the kernel functions. The kernel trick allows to apply the maximum margin algorithm to a transformed version of a non–separable dataset (feature space) via a mapping function ϕ. The related dual problem can be expressed as

$$\begin{array}{cc} max_\alpha & 2\alpha^T \mathbf{e} - \alpha^T \left(G(K) + \frac{1}{C} I_n \right) \\ st & \alpha \geq 0, \quad \alpha^T \mathbf{y} = 0 \end{array} \tag{4}$$

where $\mathbf{e}$ is the n–vector of ones, $\alpha \in \Re^n$, $G(K)$ the Gram matrix is defined by $G_{ij}(K) = [K]_{ij} y_i y_j = k(\mathbf{x}_i, \mathbf{x}_j) y_i y_j$, I_n, which is a diagonal matrix of 1 and $\alpha \geq 0$ means $\alpha_i \geq 0, \;\; i = 1, \ldots, n$. The transformation function ϕ is integrated into the definition of the Gram matrix. According to Mercer's theorem, (3) can be expressed by transforming the input data with ϕ and taking the dot product to define the kernel or taking directly any kernel and using it without knowing the function ϕ. One kind of such kernels can be the Gaussian kernel (also called radial basis functions (RBF) kernel) expressed as $K(x_i, x_j) = \phi(x_i)^T \phi(xj) = e^{-\|x_i - x_j\|^2/(2\sigma^2)}$. For such a kernel, the misclassification cost C and the kernel hyperparameter σ require optimization. The graph (A) of Figure 3 illustrates the data classification with SVM in a feature space.

Many researchers consider SVMs as one of the best classification algorithm due to its theoretical foundation based on structural risk minimization implying a better generalization performance [21]. However, SVMs may provide bad results if it is used with the wrong parameters. The usual way to find the parameters of SVMs is to scan a range of possible values of the parameters, evaluate the classifier with a data splitting such as cross–validation or a bootstrapping procedure and then select those providing the best performance. A better method to measure the generalization performance is to evaluate SVMs with the leave–one–out procedure during the grid search. These processes are expensive with respect to computation time because they require an SVM resolution

at each step. A more efficient way to choose the SVM parameters is to take advantage of the underlying theory using the bound of the leave–one–out error.

For the SVM with an RBF kernel and in the case of non–separable training data (hard margin SVM), Vapnik showed that the leave–one–out error is upper bounded by $4R^2 \|w\|^2$ (the radius margin bound) [21]. R is the radius of the smallest sphere containing all $\phi(x_i)$ and is the solution of the following optimization problem:

$$\begin{array}{rc} max_\beta & 1 - \beta^T K \beta \\ st & 0 \leq \beta_i \ i = 1, \ldots, n \\ & e^T \beta = 1 \end{array}$$

This bound of the leave–one–out error can be used to estimate the parameter σ of the RBF kernel and the soft margin parameter C. The readers are referred to [22] for a survey of the SVM error bounds estimation.

C. SVM with Asymmetrical Misclassification Cost

The SVM formulations above (Eq. 2) mean that the misclassification cost of positive and negative examples (in the case of a binary classification) are the same. This SVM formulation may be incongruous for problems with high imbalance between classes or those whose error penalty is not the same for each class.

The SVM algorithm implements natively a cost–sensitive strategy. For this purpose, two misclassification costs (C^+ for $y_i = +1$ and C^- for $y_i = -1$) are introduced. In this case, the primal formulation of the SVM is:

$$\begin{array}{rc} min_{w,b,\xi} & \langle w, w \rangle + C^+ \sum_{i \in i_+} (\xi_i^+)^2 + C^- \sum_{i \in i_-} (\xi_i^-)^2 \\ st & y_i(\langle w, \mathbf{\Phi}(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i^+, \ i \in i_+ \\ & y_i(\langle w, \mathbf{\Phi}(\mathbf{x}_i) \rangle + b) \leq -1 - \xi_i^-, \ i \in i_- \\ & \xi_i \geq 0, \ i = 1, \ldots, n \end{array} \quad (5)$$

where $i_+ = \{i | y_i = +1\}$ and $i_- = \{i | y_i = -1\}$

The corresponding dual form is:

$$\begin{array}{rc} max_\alpha & 2\alpha^T \mathbf{e} - \alpha^T \left(G(K) + \frac{1}{C^+} I_n^+ + \frac{1}{C^-} I_n^- \right) \\ st & \alpha \geq 0, \quad \alpha^T \mathbf{y} \end{array} \quad (6)$$

where $\mathbf{e}$, α and $G(K)$ has the same expression as in (Eq. 2) and I_n^+ (resp. I_n^+) is a diagonal matrix composed by 1 for $i \in i_+$ (resp. $i \in i_-$) and 0 elsewhere. It is important to highlight that for an identical misclassification cost for positive and negative examples, we obtain the formulation (Eq. 1) with (Eq. 3). Figure 3 illustrates the differences between SVM and ASVM.

Other approaches were proposed in the literature to handle imbalanced datasets with SVMs. A naive post–processing method consists of shifting the separating hyperplane far away from the positive examples. Another one, proposed in [17], applies a conformal transformation in the feature space to achieve a high separability of the training data.

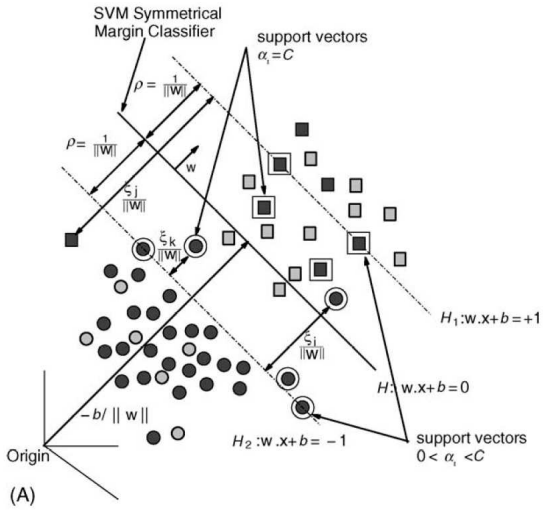

Fig. 2. Illustration of an SVM.

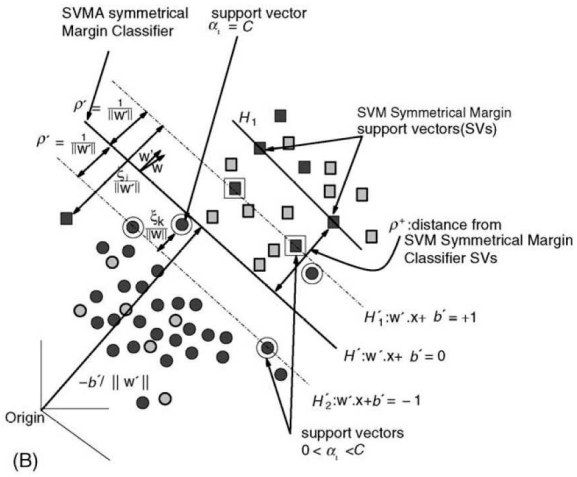

Fig. 3. Illustration of the asymmetric–margin SVM on a toy classification problem. Squares represent positive and circles negative examples; dark symbols stand for training and grey for test examples. The graph (A) shows the decision boundary induced by a conventional SVM. The graph (B) shows the new boundary obtained by introducing two cost hyperparameters respectively for positive and negative examples. Notice the two positive examples (grey squares), which are misclassified in (A) and correctly classified in (B) and also the direction of the vector w' perpendicular to the separating hyperplane in the graph (B).

D. Radius margin bound

As stated above, the radius margin bound (RMB) proposed by Vapnik is for hard–margin SVMs. To obtain the radius margin bound of the soft–margin SVM, the soft–margin should be casted into the hard margin formulation which is achieved using the following change:

$$\tilde{w} \equiv \left[\begin{array}{c} w \\ \sqrt{C}\xi \end{array} \right]$$

and set the i–th training data as

$$\left[\begin{array}{c} \phi(x_i) \\ \frac{y_i e_i}{\sqrt{C}} \end{array} \right]$$

The kernel function becomes $\tilde{K}(x_i, x_j) = K(x_i, x_j) + \delta_{ij}/C$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The new radius margin bound is $\tilde{R}^2 \|\tilde{w}\|^2$, where $\tilde{R}^2$ is the objective value of

$$\begin{array}{cc} max_\beta & 1 + \frac{1}{C} - \beta^T \left(K + \frac{1}{C}\right) \beta \\ st & 0 \leq \beta_i, i = 1, \ldots, n \\ & e^T \beta = 1 \end{array} \quad (7)$$

To solve the asymmetric–margin problem, all SVM solvers use only one value of C and balance the misclassification with weights to obtain the value of C^+ and C^-. To exploit the radius–margin for asymmetrical misclassification cost, we introduce the following relation: $C = C^+ + C^- = w_+ C + w_- C$ and $w_+ + w_- = 1$ i.e. the cost asymmetry is taken only into account during the SVM optimization problem resolution. We also use the heuristic proposed by Morik *et al.*: the potential total cost of the false positives equals the potential total cost of the false negatives i.e. the costs C^+ and C^- conform to the relation in Eq. 6 [23]. Thus, we obtain the value of $w_+ = \frac{N_-}{N}$ and $w_- = \frac{N_+}{N}$ where N_+,N_- and N are respectively the number of positive, negative and all training examples. With these weights, it is now possible to introduce a higher cost when the SVMs misclassify positive examples compared to a misclassification of negative examples.

$$\frac{C^+}{C^-} = \frac{number\ of\ negative\ training\ example}{number\ of\ positive\ training\ examples} = \frac{N^-}{N^+} \quad (8)$$

E. Gradient descent algorithm and SVM model selection

Optimization is generally concerned with the minimization (or maximization) of a function of which parameters are subject to one or more functional constraint. f is named a continuously differentiable function. A gradient descent algorithm is a method to solve an optimization problem. It looks iteratively (until a stop criterion is reached) for a direction $d_k \in \Re^n$ and $d_k \neq 0$ from a starting point $x_0 \in \Re^n$ satisfying:

$$\forall k > 1, \exists \epsilon > 0\ such\ as\ \nabla f(x_k) d_k \leq -\epsilon \|\nabla f(x_k)\| \|d_k\| \quad (9)$$

There are many ways to define the descent direction d_k. One strategy, known as the Newton method, is to use

$$d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k) \quad (10)$$

The computation of the second derivative (Hessian) of f is computationally expensive. The quasi–Newton method approximates the Hessian at each iteration. The algorithm of the hyperparameter selection using the quasi Newton method is shown below as introduced in [22].

The radius margin bound of the L2–SVMs ($p = 2$ in (3)) with the RBF kernel being continuously differentiable with respect to the parameters C and σ. Thus, the optimal parameter can be computed using the gradient descent algorithm according to the following:

$$\frac{\partial B_{RML2}}{\partial V_t} = \frac{\partial (R^2 \|w\|^2)}{\partial V_t} = \|w\|^2 \frac{\partial R^2}{\partial V_t} + R^2 \frac{\partial (\|w\|^2)}{\partial V_t} \quad (11)$$

Algorithm for model selection using the gradient descent algorithm

1. `Initialize SVM hyperparameters`
2. `Solve SVM problem using a standard SVM algorithm`
3. `Minimize the RMB according to the values of the Lagrangian multipliers with a gradient descent algorithm`
4. `Go to step 2 or stop when the minimum of the RMB is reached`

where V_t is the t–th parameter of the L2–SVM.

For $V = (C, \sigma^2)$,

$$\begin{array}{rcl} \frac{\partial \|w\|^2}{\partial C} & = & \sum_{i=1}^n \alpha_i^2 / C^2 \\ \frac{\partial \|w\|^2}{\partial \sigma^2} & = & \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \frac{\partial \tilde{k}(x_i, x_j)}{\partial \sigma^2} \\ \frac{\partial R^2}{\partial C} & = & \sum_{i=1}^n \beta_i (1 - \beta_i) / C^2 \\ \frac{\partial R^2}{\partial \sigma^2} & = & \sum_{i,j=1}^n \beta_i \beta_j \frac{\partial \tilde{k}(x_i, x_j)}{\partial \sigma^2} \\ \frac{\partial \tilde{k}(x_i, x_j)}{\partial \sigma^2} & = & \tilde{k}(x_i, x_j) \frac{\|x_i - x_j\|^2}{2\sigma^4} \end{array}$$

III. Material and Methods

This section introducea the dataset we are using for the experimentations, the software used for the extraction of features, the SVM software, the features build from the HRCT images and the learning process. The latter includes the implementation of the resampling methods, the model selection process and the details of the metric used to assess the quality of the classification.

A. Dataset

The dataset used in this work is extracted from an in–house multimedia collection of cases at the University Hospitals of Geneva (HUG). The diagnoses of each ILD cases was confirmed by a biopsy or an equivalent test (e.g. bronchoalveolar lavage, tuberculin skin test, Kveim test, ...). For each collected patient, 99 clinical parameters associated with 13 of the most frequent diagnoses of ILDs were collected from the electronic health record (EHR), describing the patient's clinical state at the time of the stay when the HRCT image series were acquired. The lung tissue patterns related to the ILD diagnosis were manually delineated in HRCT images series (1mm slice thickness, no contrast agent) by two experienced radiologists at the HUG. The distributions of the 6 most represented tissue sorts are detailed in Table I in terms of number of ROIs, volumes and number of block instances obtained as shown in Figure 1. The size of the blocks is $32 \times 32 \times 1$ pixels.

B. Software

The image processing algorithms include wavelet–based features and grey–level histograms and were implemented in Java. The classification task is carried out with libSVM implementing the SVM–L2 for binary classification [24].

TABLE I

DISTRIBUTION OF THE CLASSES IN TERMS OF ROIS, VOLUMES AND BLOCKS. THE NUMBER OF INSTANCES CORRESPONDS TO THE NUMBER OF BLOCKS.

label	ROIs	volume (liters)	blocks	patients
healthy	100	5.12 l	3043	7
emphysema	66	1.15 l	422	5
ground glass	427	4.91 l	2313	37
fibrosis	473	8.45 l	3113	38
micronodules	297	16.06 l	6133	16
consolidation	196	0.69 l	90	14
Total	1559	36.38 l	15114	87

C. Texture features

The features used to characterize the texture properties of the 6 lung tissue patterns are derived from grey–level histograms and tailored wavelet transforms (WT). The resulting feature space has a dimension of 46.

1) Grey–level histograms: Thanks to Hounsfield Units (HU), the pixel values in HRCT images corresponds univoquely to the density of the observed tissue and thus contain essential information for the characterization of the lung tissue. To encode this information, 22 histogram bins of grey–levels in the interval $[-1050; 600[$ are used as texture features. An additional feature related to the number of air pixels is computed as the number of pixel values below -1000 HU.

2) Wavelet–based features: Near affine–invariant texture features are derived from a tailored WT. A frame transform is used to ensure translation–invariant descriptions of the lung tissue patterns [10], [25]. Based on the assumption that no predominant orientations are contained in the lung tissue patterns a rotation–invariant nonseparable WT is implemented using isotropic polyharmonic B–spline scaling functions and wavelets [26], [27]. At last, an augmented scale progression is obtained using the quincunx lattice for upsampling the filters by a factor of $\sqrt{2}$ at each iteration of the WT. Within each unique subband i, the wavelet coefficients are characterized by a mixture of two Gaussians with fixed means $\mu^i_{1,2} = \mu^i$ and distinct variances $\sigma^i_{1,2}$. 24 wavelet–based features are thus generated by 8 iterations of the WT.

D. Imbalance management

The datasets we are using contain imbalance with respect to the class distribution. Three strategies were implemented to handle the imbalance. The first strategy implements the data–driven method (BAL). A random down–sampling of the majority class is carried out to obtain 50% of the positive and negative cases during the model selection. The second strategy uses the cost–sensitivity method (ASVM). The ratio of the original dataset is kept during the model selection process and the values of the cost hyperparameters were adjusted according to the imbalance rate of the positive and negative cases. The third strategy uses a combination of the resampling and cost–sensitive methods (ASVM + RES). The resampling level is based on the prevalence of the patient in the database (see Table I). If the ratio of the tissue is less than the prevalence, the examples of this class are oversampled (*consolidation*, *emphysema*, *fibrosis*, *ground glass*). If the ratio of the tissue is greater than the prevalence, the majority will be oversampled so that the ratio of positive and negative cases is equal to the prevalence (*healthy*, *micronodules*).

E. Model selection

The selection of the model was inspired by the experimental setup proposed in [28]. We have chosen 5 starting points for the gradient descent. These 5 starting points were applied to 5 random training files. The parameters obtained with an initialization point providing the least hyperparameter variance are considered. The median is taken as a new starting point and is evaluated on the whole training set by the means of a leave–one–patient–out (LOPO) procedure [29]. The final parameters are the median of those obtained from this last step. We analyze the error rate, the sensitivity, the specificity and the precision of the prediction on test sets. As we are in a multi–class classification, we use the one–against–all procedure.

F. Model comparisons

In many classification projects, the accuracy is chosen as the main performance criterion of a model. With imbalanced datasets, we have to take into account the ability of the classifier to predict the examples of each class (sensitivity, specificity and precision). These four metrics are used to measure the performance of each model. To assess the multiclass performance of the algorithms, the geometric mean is computed as follows:

$$A^{geom} = \sqrt[N_{class}]{\prod_{i=1}^{N_{class}} A_l}, \tag{12}$$

with N_{class} the number of classes and A_l the class–specific accuracies.

To evaluate the best strategy for our dataset, we carried out a McNemar test with Bonferroni correction on the prediction results. This test measures if the predictions made with 2 models are significantly different from the statistical point of view. We also use the area under the receiving operator curve (AUC) to rank the three strategies. Each strategy is assigned a score from 3 to 1 (best to worst) according to the AUC value.

IV. RESULTS

Results of the model selection and associated classification performance obtained with the various techniques are described in this section.

A. Model selection using the gradient descent

For the model selection, we used five starting points (C, σ) on five random training sets: (1,1), (5,5), (5,1), (1,5) and (10,1). Among these five initialization points, the first four converged around the same region but a few initialization

TABLE II

PERFORMANCE ON *consolidation* VS. ALL CLASSES.

Consolidation	BAL	ASVM	ASVM + RES
Error	0.02	0.02	0.02
Sensitivity	0.39	0.40	0.40
Specificity	0.99	0.99	0.99
Precision	0.16	0.16	0.16
F–measure	0.23	0.23	0.23
AUC	0.69	0.69	0.69

TABLE III

PERFORMANCE ON *emphysema* VS ALL CLASSES.

Emphysema	BAL	ASVM	ASVM + RES
Error	0.22	0.22	0.22
Sensitivity	0.45	0.46	0.46
Specificity	0.79	0.79	0.78
Precision	0.06	0.06	0.06
F–measure	0.10	0.10	0.10
AUC	0.62	0.62	0.62

points provide high variance on C. The median of these 20 intermediate parameters was used as a starting point on a LOPO cross–validation to obtain the final parameters. The algorithm converges after 7 to 18 iterations and the computation of these parameters varies from 10 minutes to 24 hours depending on the size of the resampled training data.

B. Classification performance

The Tables II, III, IV, V, VI and VII summarize the classification results obtained using the three models. AUC obtained with the various techniques are summarized in Figure 4. The best A^{geom} value of 0.752 was obtained using the ASVM+RES approach. It is followed by ASYM with $A^{geom} = 0.749$ and worst performance is obtained with BAL with $A^{geom} = 0.746$.

V. DISCUSSION

The convergence of the four initialization points to the same region indicates a consistency of the use of the RMB for SVM hyperparameters estimation. The computation time depends primarily on the size of the training set. During the model selection process i.e. LOPO cross–validation, the SVMs were run only 2 times per fold and the algorithm converged after 7 to 18 iterations.

With respect to the multiclass performance, the geometric mean ranked the third strategy (ASVM+RES) as the best choice for the lung tissue classification especially for the *fibrosis*, *ground glass* and *healthy* tissues (see for example the figure 4). The McNemar statistical test on each pair of these strategies indicates that there are no significant differences in the results obtained with the three strategies for the classification of *consolidation* and *emphysema*. The McNemar test also indicates no significant difference between ASVM and ASVM+RES for the classification of *healthy* tissue. Table VIII summarizes the ranking of the three strategies according to the value of the AUC for four tissue types.

The choice of the AUC to rank the strategies was taken because it takes into account the sensitivity and the specificity of the classifier i.e. the ratio of true positive and true negative cases. Depending on the final use of the classification models, the ranking in Table VIII may not hold anymore. For instance, if the f–measure was used to rank the three strategies as in Table VIII, the ASVM and ASVM+RES strategies would have the same rank. A ranking according to the f–measure would agree with a ranking based on AUC except for *fibrosis* where ASVM has the highest f–measure (i.e. would be ranked as the best) even if it has the lowest

TABLE IV

PERFORMANCE ON *fibrosis* VS. ALL CLASSES.

Fibrosis	BAL	ASVM	ASVM + RES
Error	0.21	0.19	0.20
Sensitivity	0.80	0.77	0.79
Specificity	0.79	0.82	0.80
Precision	0.50	0.53	0.51
F–measure	0.61	0.63	0.62
AUC	0.79	0.79	0.80

TABLE V

PERFORMANCE ON *ground glass* VS. ALL CLASSES.

Ground glass	BAL	ASVM	ASVM + RES
Error	0.46	0.46	0.45
Sensitivity	0.86	0.87	0.88
Specificity	0.48	0.48	0.5
Precision	0.23	0.23	0.24
F–measure	0.36	0.37	0.38
AUC	0.67	0.68	0.69

TABLE VI

PERFORMANCE ON *healthy* VS. ALL CLASSES.

Healthy	BAL	ASVM	ASVM + RES
Error	0.24	0.23	0.23
Sensitivity	0.95	0.95	0.95
Specificity	0.71	0.72	0.72
Precision	0.45	0.46	0.46
F–measure	0.61	0.62	0.62
AUC	0.83	0.83	0.84

TABLE VII

PERFORMANCE ON *micronodules* VS. ALL CLASSES.

Micronodules	BAL	ASVM	ASVM + RES
Error	0.31	0.32	0.30
Sensitivity	0.55	0.53	0.45
Specificity	0.79	0.79	0.87
Precision	0.64	0.63	0.69
F–measure	0.59	0.57	0.54
AUC	0.67	0.66	0.66

TABLE VIII

RANKING OF THE THREE STRATEGIES ACCORDING TO THE AUC VALUES.

	BAL	ASVM	ASVM+RES
fibrosis	1	2	3
ground glass	1	2	3
healthy	1	2	3
micronodules	3	2	1
Total ranking	6	8	10

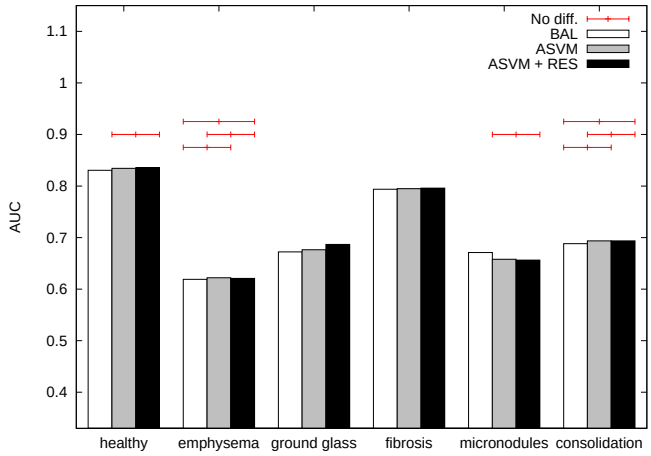

Fig. 4. AUC values obtained with the various techniques. The horizontal bars at the top of the histogram indicate if the two results has no significant difference according to the McNemar statistical test.

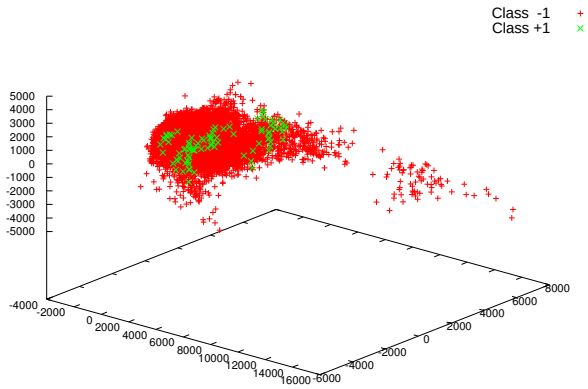

Fig. 5. Projection of the *consolidation* vs all dataset on the first three principal component axes. The *consolidation* examples are labelled as +1.

sensitivity.

The BAL strategy outperforms than other strategies for the classification of *micronodules*. The asymmetric–margin of the SVM altered the sensitivity of the model. The addition of synthetic majority examples combined with an asymmetric–margin SVM has further deteriorated the sensitivity of the model but improved the precision and the specificity of the classifier. A possible explanation of this phenomenon is the existence of outliers in the *micronodules* examples, which were misclassified with the ASVM and ASVM+RES and thus decreasing the sensitivity and the precision.

The results obtained in the classification of *healthy* tissue is of particular interest in clinical practice where the three models have sensitivity equal to 95%. Indeed, an ASVM model for the classification of healthy tissue can be used to detect abnormal tissue types in HRCT.

Table II highlights the question of using the accuracy as a performance metric in the classification of imbalanced data sets. According to this table, the accuracy is around 98% but the algorithm correctly classified 99% of the negative examples while the sensitivity and precision on positive examples are low.

The precision for the classification of *consolidation*, *emphysema* and *ground glass* are very low and more investigations are needed to improve the results. This is probably due to the characteristics of these tissues: they are not very specific according to a discrimination measure of these tissues against the rest. The latter was carried out with the Rayleigh quotient (ratio of the between–class variance and the within–class variance) on the dataset projected into the principal component axis. The figure 5 highlights the high disparity of the consolidation in the principal components space.

Many investigations will be carried out in the future to improve the classification of the three lung tissues cited in the previous paragraph. Increasing the number of examples of these tissues may provide better improvement of the classification results. The *consolidation* tissues, for examples, are from younger patients compared to the other patients in the database. Another avenue for future investigation is to play with the oversampling ratio. The computational speed of the gradient descent method (compared to a fine–grained grid–search) and the stability of the model will allow us to carry out more experimentations with respect to the oversampling ratio of the minority classes. The addition of clinical features may also increase the performance of the classification. The use of kernel–based algorithm to transform the input space into a linearly separable dataset did not provide good results and it is possible that the RBF kernel is not well-suited for the classification of these tissues. Other approaches allowing the selection of the appropriate kernel also figures in the list of future works [30].

VI. Conclusions

We presented in this paper the effectiveness of asymmetric–margin SVMs for imbalanced lung tissue classification. The introduction of prior knowledge of the prevalence of the patients in the database to correct the ratio of the examples improved the results with the algorithm. Artificial cases were created according to the k–means algorithm. The conventional SVM was only better in the classification of *micronodules* due to the presence of outliers in the examples. While the results obtained with ASVM for the classification of *fibrosis* and *healthy* tissues are satisfactory, more investigations are needed for the classification of *consolidation*, *emphysema*, *ground glass* and also *micronodules*. Increasing the number of cases, varying the oversampling ratio, addition of clinical features and selection of appropriate kernel for each classification are the most important for future investigations.

Acknowledgment

This work was supported by the Swiss National Science Foundation (FNS) with grant 200020–118638/1 and

the equalization fund of Geneva University Hospitals and University of Geneva (grant 05–I–13 and 05–9–II)

REFERENCES

[1] K. R. Flaherty, T. E. King, J. Ganesh Raghu, J. P. Lynch III, T. V. Colby, W. D. Travis, B. H. Gross, E. A. Kazerooni, G. B. Toews, Q. Long, S. Murray, V. N. Lama, S. E. Gay, and F. J. Martinez, "Idiopathic interstitial pneumonia: What is the effect of a multidisciplinary approach to diagnosis?" *American Journal of Respiratory and Critical Care Medicine*, vol. 170, pp. 904–910, July 2004.

[2] W. R. Webb, N. L. Müller, and D. P. Naidich, Eds., *High–Resolution CT of the Lung*. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2001.

[3] S. Delorme, M.-A. Keller-Reichenbecher, I. Zuna, W. Schlegel, and G. Van Kaick, "Usual interstitial pneumonia: Quantitative assessment of high–resolution computed tomography findings by computer–assisted texture–based image analysis," *Investigative Radiology*, vol. 32, no. 9, pp. 566–574, September 1997.

[4] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "ASSERT: A physician–in–the–loop content–based retrieval system for HRCT image databases," *Computer Vision and Image Understanding (special issue on content–based access for image and video libraries)*, vol. 75, no. 1/2, pp. 111–132, July/August 1999.

[5] R. Uppaluri, E. A. Hoffman, M. Sonka, G. W. Hunninghake, and G. McLennan, "Interstitial lung disease: A quantitative study using the adaptive multiple feature method," *American Journal of Respiratory and Critical Care Medicine*, vol. 159, no. 2, pp. 519–525, February 1999.

[6] I. C. Sluimer, P. F. van Waes, M. A. Viergever, and B. van Ginneken, "Computer–aided diagnosis in high resolution CT of the lungs," *Medical Physics*, vol. 30, no. 12, pp. 3081–3090, December 2003. [Online]. Available: http://link.aip.org/link/?MPH/30/3081/1

[7] F. Chabat, G.-Z. Yang, and D. M. Hansell, "Obstructive lung diseases: Texture classification for differentiation at CT," *Radiology*, vol. 228, no. 3, pp. 871–877, September 2003. [Online]. Available: http://radiology.rsnajnls.org/cgi/content/abstract/228/3/871

[8] Y. Uchiyama, S. Katsuragawa, H. Abe, J. Shiraishi, F. Li, Q. Li, C.-T. Zhang, K. Suzuki, and K. Doi, "Quantitative computerized analysis of diffuse lung disease in high–resolution computed tomography," *Medical Physics*, vol. 30, no. 9, pp. 2440–2454, September 2003.

[9] T. Zrimec and J. S. J. Wong, "Improving computer aided disease detection using knowledge of disease appearance," in *MEDINFO 2007. Proceedings of the 12th World Congress on Health (Medical) Informatics*, vol. 129. IOS Press, August 2007, pp. 1324–1328.

[10] A. Depeursinge, D. Sage, A. Hidki, A. Platon, P.-A. Poletti, M. Unser, and H. Müller, "Lung tissue classification using Wavelet frames," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. Lyon, France: IEEE Computer Society, August 2007, pp. 6259–6262.

[11] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis Journal*, vol. 6, no. 5, November 2002.

[12] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one–sided selection," in *Proceedings of the 14th International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[14] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection," *Artificial Intelligence in Medicine*, vol. 37, no. 1, pp. 7–18, May 2006. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2005.03.002

[15] H. Han, W. Wang, and B. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing (ICIC)*, 2005, pp. 878–887.

[16] J. V. Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *ICML*, 2007, pp. 935–942.

[17] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the International Joint Conference on AI*, 1999, pp. 55–60.

[18] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *In ICML 2003 Workshop on Learning from Imbalanced Data Sets*, 2003, pp. 49–56.

[19] Y.-J. Lee and O. L. Mangasarian, "Ssvm: A smooth support vector machine for classification," *Computational optimization and applications*, vol. 20, no. 1, pp. 5–22, 2001.

[20] O. Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 19, pp. 1155–1178, 2007.

[21] V. Vapnik, *Statistical learning theory*. Wiley, New York, NY, 1998.

[22] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 131–159, 2002.

[23] K. Morik, M. Imhoff, P. Brockhausen, T. Joachims, and U. Gather, "Knowledge discovery and knowledge validation in intensive care," *Artificial Intelligence in Medicine*, vol. 19, no. 3, pp. 225–249, 2000.

[24] K.-M. Chung, W.-C. Kao, C.-L. Sun, L.-L. Wang, and C.-J. Lin, "Radius margin bounds for support vector machines with the rbf kernel," *Neural Computation*, vol. 15, no. 11, pp. 2643–2681, 2003.

[25] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Transactions on Image Processing*, vol. 4, no. 11, pp. 1549–1560, November 1995.

[26] A. Depeursinge, D. Van De Ville, M. Unser, and H. Müller, "Lung tissue analysis using isotropic polyharmonic B–spline wavelets," in *MICCAI 2008 Workshop on Pulmonary Image Analysis*, New York, USA, September 2008, pp. 125–134.

[27] D. Van De Ville, T. Blu, and M. Unser, "Isotropic polyharmonic B–Splines: Scaling functions and wavelets," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1798–1813, November 2005.

[28] G. Rtsch, T. Onoda, K.-R. Mller, and T. O. Gmd, "Soft margins for adaboost," *Journal of Machine Learning*, vol. 42, no. 3, pp. 287–320, 1998.

[29] M. Dundar, G. Fung, L. Bogoni, M. Macari, A. Megibow, and B. Rao, "A methodology for training and validating a CAD system and potential pitfalls," *International Congress Series*, vol. 1268, pp. 1010–1014, June 2004, cARS 2004 – Computer Assisted Radiology and Surgery. Proceedings of the 18th International Congress and Exhibition. [Online]. Available: http://www.sciencedirect.com/science/article/B7581-4CHRSVD-6S/2/06d1476fa7e0028d30aa5db70037f836

[30] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *ICML*, 2007, pp. 775–782.