35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Pruned Ensemble Classifier for Effective Breast Thermogram Analysis

Bartosz Krawczyk and Gerald Schaefer

Abstract— **Thermal infrared imaging has been shown to be useful for diagnosing breast cancer, since it is able to detect small tumors and hence can lead to earlier diagnosis. In this paper, we present a computer-aided diagnosis approach for analysing breast thermograms. We extract image features that describe bilateral differences of the breast regions in the thermogram, and then feed these features to an ensemble classifier. For the classification, we present an extension to the Under-Sampling Balanced Ensemble (USBE) algorithm. USBE addresses the problem of imbalanced class distribution that is common in medical decision making by training different classifiers on different subspaces, where each subspace is created so as to resemble a balanced classification problem. To combine the individual classifiers, we use a neural fuser based on discriminants and apply a classifier selection procedure based on a pairwise double-fault diversity measure to discard irrelevant and similar classifiers. We demonstrate that our approach works well, and that it statistically outperforms various other ensemble approaches including the original USBE algorithm.**

I. Introduction

Thermography uses a camera with sensitivities in the thermal infrared to capture the temperature distribution of the human body or parts thereof. In contrast to other modalities such as mammography, it is a non-invasive, non-contact, passive and radiation-free technique. It is well known that the radiance from human skin is an exponential function of the surface temperature which in turn is influenced by the level of blood perfusion in the skin. Thermal imaging is hence well suited to pick up changes in blood perfusion which might occur due to inflammation, angiogenesis or other causes [1]. Thermography has also been shown to be well suited for the task of detecting breast cancer [2], [3]. Here, thermography has advantages in particular when the tumor is in its early stages or in dense tissue. Early detection is crucial as it provides significantly higher chances of survival [4] and in this respect infrared imaging can outperform the standard method of mammography. While mammography can detect tumors only once they exceed a certain size, even small tumors can be identified using thermal infrared imaging due to the high metabolic activity of cancer cells which leads to an increase in local temperature that can be picked up in the infrared [5].

In this paper, we present a computer-aided diagnosis approach for analysing breast thermograms. For this, we extract a set of image features from the thermograms that describe bilateral differences between the two breast regions (since the presence of tumors will typically lead to asymmetries between the temperature distributions of the two sides). These features are then utilised in a pattern classification stage for which we employ an ensemble classifier.

In particular, we present Prunded Under-Sampling Balanced Ensemble (PUSBE), a classification method based on our earlier work in [6], [7] for effective classification of breast thermogram features. Our approach addresses the problem of class imbalance that is frequently encountered in medical datasets due to the relatively low number of malignant cases compared to benign ones. We achieve this by constructing subspaces from balanced subsets of the training data and then train a separate classifier for each generated subspace. To avoid using similar predictors we then perform a classifier selection and pruning stage to discard redundant base classifiers. For this, we utilise a pairwise double-fault diversity measure to ensure that classifiers in the pool are mutually complementary. We combine the remaining classifiers using a trained fusion algorithm based on discriminants and implemented as a neural network. Experimental results obtained on a large dataset of breast thermograms demonstrate that our approach achieves high classification performance without sacrificing sensitivity while statistically outperforming other ensemble classification approaches.

II. Breast thermogram image features

As has been shown, an effective approach to detect breast cancer based on thermograms is to study the symmetry between the left and right breast regions [8]. In the case of cancer presence, the tumor will recruit blood vessels resulting in hot spots and a change in vascular pattern, and hence an asymmetry between the temperature distributions of the two breasts. On the other hand, symmetry typically identifies healthy subjects.

We follow this approach and extract image features that describe bilateral differences between the areas of the left and right breasts extracted from frontal view thermograms. We employ the same image features that were used in [9], namely:

- Basic statistical features: mean, standard deviation, median, 90-percentile;
- Moment features: centre of gravity, distance between moment centre and geometrical centre;
- Histogram features: cross-correlation between histograms; maximum, number of non-empty bins, number

B. Krawczyk (bartosz.krawczyk@pwr.wroc.pl) is with the Department of Systems and Computer Networks, Wrocław University of Technology, Wrocław, Poland.

G. Schaefer (gerald.schaefer@ieee.org) is with the Department of Computer Science, Loughborough University, Loughborough, U.K

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

of zero-crossings, energy and difference of positive and negative parts of difference histogram;

- Cross co-occurrence matrix [10] features: homogeneity, energy, contrast, symmetry and the first 4 moments of the matrix;
- Mutual information between the two temperature distributions;
- Fourier spectrum features: the difference maximum and distance of this maximum from the centre.

Each breast thermogram is thus described by 4 basic statistical features, 4 moment features, 8 histogram features, 8 cross co-occurrence features, mutual information and 2 Fourier descriptors. We further apply a Laplacian filter to enhance the contrast and calculate another subset of features (the 8 cross co-occurrence features together with mutual information and the 2 Fourier descriptors) from the resulting images, and consequently end up with a total of 38 features which describe the asymmetry between the two sides and which form the basis for the following pattern classification stage.

III. Ensemble classification

Assume that we have L classifiers $\Psi^{(1)}$, $\Psi^{(2)}$, ..., $\Psi^{(L)}$. For a given object $x \in \mathcal{X}$, each individual classifier decides for class $i \in \mathcal{M} = \{1, \dots, M\}$ based on the values of discriminants. Let $F^{(l)}(i, x)$ denote a function that is assigned to class i for a given value of x, and that is used by the l-th classifier $\Psi^{(l)}$. The combined classifier Ψ makes a decision based on [11]

$$\Psi(x) = i \quad if \quad \hat{F}(i, x) = \max_{k \in M} \hat{F}(k, x), \tag{1}$$

where

$$\hat{F}(i, x) = \sum_{l=1}^{L} w^{(l)}(i) F^{(l)}(i, x) \quad \text{and} \quad \sum_{i=1}^{L} w^{(l)}(i) = 1. \tag{2}$$

The weights can be set dependent on the classifier and class number: weight $w^{(l)}(i)$ is assigned to the l-th classifier and the i-th class, and given classifier weights assigned to different classes may differ.

IV. Imbalanced classification

A data set is imbalanced if the classification categories are not (approximately) equally represented. Especially in medical decision making, data sets are often predominantly composed of "normal" or benign examples with only a small percentage of "abnormal" or malignant cases.

While the performance and quality of classification algorithms is usually evaluated using predictive accuracy, this is not appropriate when the data is imbalanced as disproportions in the number of objects between the classes may lead to severe deterioration of the classification accuracy. Consequently, the decision boundary may get biased towards the majority class, leading to poor recognition of the minority class and hence poor sensitivity.

Various approaches have been proposed to address this problem. Among the most popular ones are oversampling [12], which introduces artificial objects into the dataset to counter the unfavourable sample distribution, and cost-sensitive classification [13] where a cost matrix is defined to associate misclassification with costs and classification is tuned to minimise overall costs.

V. Pruned Under-Sampling Balanced Ensemble

In our approach, we perform neither oversampling (as it might lead to a class distribution shift) nor cost-sensitive classification (as performance relies heavily on the correct specification of the cost matrix). Our method is based on our earlier work in [6], [7] where we have presented Under-Sampling Balanced Ensemble (USBE) as an effective method for imbalanced classification. USBE is based on the idea of object space partitioning where each classifier is trained on a different subspace and constructed so as to counter the original imbalance in the dataset. In this paper, we add a classifier selection step to discard similar classifiers which do not contribute to the ensemble under consideration but increase its overall complexity.

Our Pruned USBE (PUSBE) approach proceeds in four main steps:

1) Creation of a number of balanced subspaces consisting of minority class and under-sampled majority class objects.
2) Construction of a pool of classifiers by training a single classifier on each of the subspaces. Optionally, a feature selection algorithm [14] can be employed which is applied independently for each of the subspaces/classifiers.
3) Diversity-based pruning of a pool of classifiers to select complementary models for the committee.
4) Fusion of outputs of the remaining classifiers.

In the following, we describe these stages in more detail.

A. Space partitioning

For imbalanced datasets, typically the majority class is identified well (as it has sufficient training instances to learn from) while classification for the minority class is often poor. In USBE this problem is addressed based on object space division and proceeds in two steps:

1) Creation of a number of subspaces.
2) Construction of a pool of classifiers $\Pi^{\Psi} = \{\Psi_1, \Psi_2, ..., \Psi_L\}$ by training single classifiers on each of the subspaces.

Space partitioning is employed to balance the unfavourable class distribution using a random undersampling method. Each of the newly created subspaces contains a smaller number of objects, randomly drawn from the dataset so that the number of objects in each of the subspaces is equal for both classes.

B. Feature selection

Feature selection is employed independently for each of the chosen subspaces. Therefore, in each of the subspaces the derived feature subsets may vary, leading to an increased overall diversity of the pool of classifiers, and consequently to a better ensemble. We employ the fast correlation-based feature filter (FCBF) [14], due to its typically good performance and low computational complexity. In FCBF, the relations between features-classes and between pairs of features are considered. The algorithm proceeds at two levels. First, a ranking algorithm using the symmetric uncertainty coefficient (SUC) index is used to estimate class-feature relevance, and a threshold coefficient established to select predominant features. In the second part, features that are redundant to the predominant features are removed.

C. Ensemble pruning

Different base classifiers will have different areas of competence and hence may provide different contributions to the committee. Therefore, careful classifier selection should be conducted in order to choose the most valuable committee members. There are several ways how such an ensemble pruning procedure can be implemented. One of the most popular criteria is to employ an ensemble diversity measure to select classifiers that are as different as possible from each other. This is motivated by the fact, that adding similar classifiers to the committee does not improve its quality but only increases its complexity. On the other hand, diverse models might be mutually supplementary and hence allow to exploit different areas of competence.

In our PUSBE approach, we employ a pairwise double-fault diversity measure [15] to select classifiers and prune the ensemble. The diversity measure is based on the idea that it is more important to know when simultaneous misclassifications occur than when both classifiers are correct. This is also well aligned with the problem of imbalanced classification, since the main priority here is to minimise the number of misclassifications of the minority class.

Given two base classifiers h_i and h_j let $n(a,b)$ denote the number of training objects for which the output of these classifiers is a and b respectively. The double-fault diversity measure can then be calculated as

$$DIV_{DF}(h_i,h_j) = \frac{n(-1,-1)}{n(1,1)+n(-1,1)+n(1,-1)+n(-1,1)}. \quad (3)$$

Diversity for an ensemble of L base classifiers is then calculated by averaging the measure over all classifier pairs in the ensemble

$$DIV_{DF}(\Psi) = \frac{2}{NL(L-1)} \sum_{j=1}^{L} \sum_{k=j+1}^{L} n_{j,k}(-1,-1), \quad (4)$$

where N is the number of training samples. The established diversity measure is in the interval $[0,1]$, where 1 corresponds to a set of identical classifiers and 0 to the highest possible diversity respectively.

D. Classifier fusion

Classifier fusion is an important aspect of classifier ensembles, and the choice of fusion method, which is responsible for the collective decision making process, is hence crucial [16]. In our approach, we use a neural network as a trained fuser for the classifier ensemble [11]. For this approach, all simple classifiers must give decisions based on the values of discriminant functions.

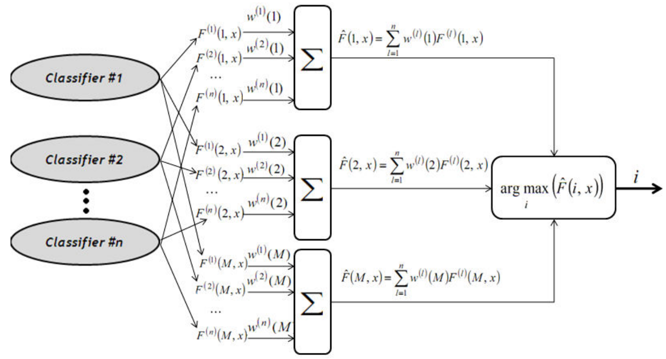

Fig. 1. Classifier fuser implemented as a one-layer neural network.

Based on a training process, the fuser needs to identify $W = \{W_1, W_2, \ldots, W_L\}$ which defines the weights assigned to each classifier and each of the M classes

$$W_a = \left[w^{(l)}(1),\ w^{(l)}(2), \ldots, w^{(l)}(M)\right]^T. \quad (5)$$

In PUSBE, we employ a neural network as a trained classifier fusion approach, illustrated in Figure 1. One perceptron fuser is constructed for each of the classes under consideration. Once trained (we employ the Quickprop algorithm in our implementation), the input weights established during the learning process are then the weights assigned to each of the base classifiers.

VI. Experimental results

For evaluation, we used a dataset of 146 thermograms (the same dataset used in [10], [9], [6], [7]) of which 29 cases have been confirmed as malignant whereas the other 117 cases were benign, and which hence clearly represents an imbalanced classification problem. For all thermograms, we extracted the 38 features described in Section II.

As base classifier, we employed a support vector machine (SVM) [17] with a Gaussian RBF kernel, and performed classifier tuning [18] to obtain optimal parameters (which were $\sigma = 0.1$ and $C = 10$). We then performed classification using the presented PUSBE approach. Classifier selection was achieved by an exhaustive search over all possible permutations of committee members to minimise the diversity measure function. The initial ensemble consisted of 7–9 individual classifiers (depending on the fold of the employed cross validation) of which 3–6 remained after the pruning stage.

For comparison, we also performed classification using several state-of-the-art ensembles dedicated to imbalanced classification, namely SMOTEBagging [19], SMOTEBoost [20], IIVotes [21] and EasyEnsemble [22], as well as the original USBE approach from [7]. For all models we used the same base classifier models as employed for PUSBE for an objective comparison.

Results, based on a combined 5 x 2 CV F test of statistical significance [23], are given in Table I for all classifier ensem-

TABLE I

CLASSIFICATION RESULTS.

classifier	sensitivity	specificity	accuracy	sensitivity statistically better than
SMOTEBagging	77.35	90.50	87.89	-
SMOTEBoost	79.03	91.00	88.62	SMOTEBagging
IIVotes	79.56	91.89	89.44	SMOTEBagging
EasyEnsemble	80.02	90.17	88.22	SMOTEBagging, SMOTEBoost
USBE	80.35	90.15	88.21	SMOTEBagging, SMOTEBoost, IIVotes
PUSBE	81.37	90.59	88.76	ALL OTHER METHODS

bles. For each approach, we report sensitivity (i.e. probability that a case identified as malignant is indeed malignant), specificity (i.e. probability that a case identified as benign is indeed benign) and overall classification accuracy (i.e. percentage of correctly classified patterns), and also state compared to which other approaches a method was found to work statistically better.

From Table I we can see that while our proposed PUSBE method does not outperform the other ensembles in terms of overall classification accuracy, clearly improved sensitivity results are achieved. The achieved sensitivity is 81.37%, the highest of all approaches, which coupled with a specificity of 90.59% gives very good classification performance on this challenging dataset. In fact, PUSBE gives statistically better sensitivity compared to all other approaches including the original USBE algorithm, which confirms that our approach is able to correctly identify the more important (i.e. malignant) cases well.

VII. CONCLUSIONS

In this paper, we have presented an effective approach to computer-aided diagnosis of breast thermograms. Based on a set of asymmetry features extracted from the images, we employ a multiple classifier system that is able to appropriately deal with imbalanced datasets. In particular, we train the base classifiers on different object subspaces where in each subspace a balance between the classes is maintained. Redundant classifiers are then removed based on a pairwise double-fault diversity measure, and the remaining classifiers combined using a neural network fuser. Our approach is shown to provide very good classification results on a challenging dataset of about 150 thermograms, and is further demonstrated to outperform various other state-of-the-art ensemble classification methods.

REFERENCES

[1] B. Jones, "A reappraisal of infrared thermal image analysis for medicine," *IEEE Trans. Medical Imaging*, vol. 17, no. 6, pp. 1019–1027, 1998.

[2] N. Anbar, L. Milescu, A. Naumov, C. Brown, T. Button, C. Carly, and K. AlDulaimi, "Detection of cancerous breasts by dynamic area telethermometry," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 5, pp. 80–91, 2001.

[3] J. Head, F. Wang, C. Lipari, and R. Elliott, "The important role of infrared imaging in breast cancer," *IEEE Engineering in Medicine and Biology Magazine*, vol. 19, pp. 52–57, 2000.

[4] M. Gautherie, "Thermobiological assessment of benign and maligant breast diseases," *Am. J. Obstet. Gynecol.*, vol. 147, no. 8, pp. 861–869, 1983.

[5] J. Keyserlingk, P. Ahlgren, E. Yu, N. Belliveau, and M. Yassa, "Functional infrared imaging of the breast," *IEEE Engineering in Medicine and Biology Magazine*, vol. 19, no. 3, pp. 30–41, 2000.

[6] B. Krawczyk and G. Schaefer, "Evolutionary multiple classifier system based on space partitioning for breast thermogram analysis," in *16th Online World Conference on Soft Computing in Industrial Applications*, 2011.

[7] ——, "Effective multiple classier systems for breast thermogram analysis," in *21st Int. Conference on Pattern Recognition*, 2012, pp. 3345–3348.

[8] H. Qi, W. E. Snyder, J. F. Head, and R. L. Elliott, "Detecting breast cancer from infrared images by asymmetry analysis," in *22nd IEEE Int. Conference on Engineering in Medicine and Biology*, 2000.

[9] G. Schaefer, M. Zavisek, and T. Nakashima, "Thermography based breast cancer analysis using statistical features and fuzzy classification," *Pattern Recognition*, vol. 42, no. 6, pp. 1133–1137, 2009.

[10] M. Zavisek and A. Drastich, "Thermogram classification in breast cancer detection," in *3rd European Medical and Biological Engineering Conference*, 2005, pp. 1727–1983.

[11] M. Woźniak and M. Zmyslony, "Designing combining classifier with trained fuser - analytical and experimental evaluation," *Neural Network World*, vol. 20, no. 7, pp. 925–934, 2010.

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

[13] B. Krawczyk, M. Wozniak, and G. Schaefer, "Improving minority class prediction using cost-sensitive ensembles," in *16th Online World Conference on Soft Computing in Industrial Applications*, 2011.

[14] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, pp. 1205–1224, 2004.

[15] G. Giancinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image Vision and Computing Journal*, vol. 19, pp. 699–707, 2001.

[16] B. Krawzcyk and G. Schaefer, "Ensemble fusion methods for medical data classification," in *11th Symposium on Neural Network Applications in Electrical Engineering*, 2012.

[17] V. Vapnik, *Statistical Learning Theory*. Willey, 1998.

[18] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "Kernlab, an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, 2004.

[19] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *IEEE Symp. Comput. Intell. Data Mining*, 2009, pp. 324–331.

[20] N. Chawla, A. Lazarevic, L. Hall, and K. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," in *Lecture Notes in Artificial Intelligence*, vol. 2838, 2003, pp. 107–119.

[21] J. Blaszczynski, M. Deckert, J. Stefanowski, and S. Wilk, "Integrating selective pre-processing of imbalanced data with ivotes ensemble," ser. Lecture Notes in Computer Science, 2010, vol. 6086 LNAI, pp. 148–157.

[22] X. Liu, J. Wu, and Z. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.

[23] E. Alpaydin, "Combined 5 x 2 cv f test for comparing supervised classification learning algorithms," *Neural Computation*, vol. 11, no. 8, pp. 1885–1892, 1999.