Effective Imbalanced Classification of Breast Thermogram Features

Bartosz Krawczyk[1] and Gerald Schaefer[2(✉)]

[1] Department of Systems and Computer Networks,
Wroclaw University of Technology, Wrocław, Poland
[2] Department of Computer Science, Loughborough University,
Loughborough, UK
gerald.schaefer@ieee.org

Abstract. Breast cancer is the most commonly occurring form of cancer in women, and can be diagnosed using various imaging modalities including thermography. In this paper, we present an approach to analysing breast thermograms based on statistical image features and an effective ensemble method for imbalanced classification problems. We extract a series of features from the images to arrive at indications of asymmetry between left and right breast regions. These then form the input to a classification stage for which we develop a dedicated multiple classifier system that employs neural networks or support vector machines as base classifiers, trains base classifiers on balanced subsets of the training data to address the class imbalance that is typically inherent in medical decision making problems, and fuses the decisions using a neural network combined with a fuzzy diversity measure to remove individual classifiers from the ensemble and to enhance prediction performance. Experimental results, on a large dataset of about 150 breast thermograms, confirm our approach to provide excellent classification performance and to outperform other classifier ensembles designed for imbalanced datasets.

Keywords: Breast cancer · Thermography · Pattern classification · Imbalanced classification · Multiple classifier system

1 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 [14]. Thermography has also been shown to be well suited for the task of detecting breast cancer [3,13]. 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

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 535–544, 2015.
DOI: 10.1007/978-3-319-19941-2_51

survival [12] 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 [16].

In our approach, which is a continuation of the work presented in [18], we therefore derive a set of image features that describe possible asymmetries between the bilateral breast regions to capture this effect. These features are then used in a pattern classification stage for which we develop a multiple classifier system (MCS). In particular, we employ neural networks or support vector machines as base classifiers, and, importantly, address the problem of class imbalance, that often occurs in medical data analysis, by training the individual classifiers on balanced data subsets, thus eliminating any unfavourable class distribution. The base classifiers are then combined using an fuser implemented as a one-layer perceptron neural network. Finally, we remove redundant classifiers through an ensemble diversity measure based on fuzziness using an energy approach. Experimental results, on a dataset of about 150 breast thermograms, confirm that our proposed approach works well and gives excellent classification performance. We furthermore show it to statistically outperform not only canonical classifiers but also recent classifier ensembles that are also dedicated to imbalanced classification.

2 Background

Several computer aided diagnostic (CAD) approaches to analysing breast thermograms have been presented in the literature. In [23], an attempt based on asymmetry analysis is presented where, following segmentation based on edge detection and the Hough transform, Bezier histograms are generated and compared to identify cancer cases. In [25], some basic statistical features are extracted and passed to a complementary learning fuzzy neural network (CLFNN) for diagnosis. Reference [26] proposes morphological analysis of "localised temperature increase" amplitudes in thermograms to detect tumors. A series of image features from the breast regions (the same features that we employ in this paper) are extracted in [24] and subsequently analysed by a fuzzy classification method, while [31] uses the same feature set in conjunction with a neural network classifier. The approach in [6] is based on transforming the thermogram into a representation derived from independent component analysis, thresholding and correlating the obtained channels to locate tumor areas. In [1], texture features and support vector machine classifiers are employed, while in [22] wavelet and texture descriptors are used in combination with several classification algorithms.

3 Image (A)symmetry 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 [23].

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 [24] (for a more extensive discussion of them, see there), 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 of zero-crossings, energy and difference of positive and negative parts of difference histogram;
- Cross co-occurrence matrix [31] 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.

4 Imbalanced Pattern Classification Ensemble

In our approach, we employ an ensemble classifier, i.e. perform classification not based on a single algorithm but based on a joint decision of a committee of classifiers [20]. This way, we are able to exploit the strengths of different base classifiers while eliminating their weaknesses, thus leading to more robust and typically better classification performance.

Given a pool of N classifiers $\Psi^{(1)}, \Psi^{(2)}, \ldots, \Psi^{(N)}$, for a given feature vector x, each of the individual classifiers makes a decision with respect to class $i \in \mathbf{M} = \{1, \ldots, M\}$. The classifier ensemble $\bar{\Psi}$ then makes a combined decision based on

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

where

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

and $F^{(l)}(i,x)$ is a discriminant function for the i-th class and object x used by the l-th classifier. The weights $w^{(l)}$ here play a crucial role for the performance of the ensemble and can be assigned either statically and through training.

In the following, we detail the components of our classifier ensemble.

4.1 Base Classifiers

While in principle any classification approach can serve as base classifier, in our approach we build ensembles of neural network (NN) or support vector machine (SVM) classifiers. For the NN classifier [4], we use the Quickprop algorithm for training, and set the number of hidden neurons to half the sum of input and output neurons. For the SVM classifier [27], we employ a Gaussian RBF kernel, and perform classifier tuning [15] to obtain optimal parameters.

4.2 Imbalanced Classification

In medical diagnosis, there are typically far fewer malignant cases than there are benign ones, and consequently conventional classification approaches often suffer from low sensitivity due to the skewed class distribution. Class imbalance can be addressed in several ways, including oversampling of the minority class [9] and cost-sensitive classification [19].

In our approach, we revert to neither of these, but employ our earlier method from [18] which is based on the principle of object space partitioning to train individual classifiers on balanced subsets of the training data.

In particular, we create a number of subspaces using a random undersampling method. Each of the subspaces contains a smaller number of objects, randomly drawn from the dataset, so that the number of objects from each of the classes are equal. Objects of the majority class are randomly sampled and removed from the training set. Subspaces are then created as long as there are objects in the majority set. Each subspace forms the basis of one of the classifiers; that is, each base classifier is trained on a different (balanced) training subset, hence leading to a heterogeneous ensemble that addresses class imbalance.

To boost recognition performance, a feature selection step is performed. For this purpose, we utilise the fast correlation-based feature filter (FCBF) [30]. 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 index is employed to estimate class-feature relevance, and a threshold established to select predominant features. In the second part, features redundant to the predominant features are removed. Since feature selection is applied separately for each subspace, and hence each classifier, this step also enhances the heterogeneity of the ensemble.

4.3 Ensemble Diversity

Different base classifiers will have different areas of competence and hence may provide different contributions to the committee. Careful classifier selection should be hence conducted in order to choose the most valuable individual

models. Therefore, in this paper and in contrast to [18], we employ a classifier ensemble diversity measure for this purpose. For this purpose, we extend the energy-based fuzzy diversity measure introduced in [17] for one-class classification problems to multi-class classification.

The proposed energy approach provides an effective measure of fuzziness. It uses a threshold $\lambda \in [0,1]$ whose role is to filter insignificant degrees of membership, that may otherwise lead to lowering the stability of the measure. Given N base classifiers in the pool, out of which S correctly classify a given training object x_j, one can define a fuzzy membership function $\mu_{x_j} = \frac{S}{N}$ for the given object, with $0 \leq \mu_{x_j} \leq 1$.

Based on this, the employed energy measure is calculated as

$$DIV = \int_X \sum_{i=1}^{N} f_\lambda(x) dx, \tag{3}$$

where

$$f_\lambda(x) = f(x) \Leftrightarrow \frac{\sum_{k=1}^{N} \delta(\Psi_{ik}^M(x), \Psi^*(x))}{N} > \lambda, \tag{4}$$

and $\Psi^*(x)$ denotes a classifier correctly classifying object x, and $f(x) : [0,1] \rightarrow R_+$ is an increasing function in interval $[0,1]$ for $f(0) = 0$.

The derived measure gives an indication of the diversity of the entire classifier committee in the range $[0,1]$, where 0 corresponds to an ensemble of identical classifiers and 1 to the highest possible diversity respectively.

We perform diversity-based classifier selection through an exhaustive search over all possible combinations of committee members, and selecting the ensemble that yields maximal diversity.

4.4 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 by determining the weights in Eq. (2), is hence crucial. Instead of traditional approaches such as majority voting or static weight assignment, in this paper we utilise a dynamic approach to combine the outputs of base classifiers created on different object subspaces. In particular, we employ a trained fuser which, although taking longer to achieve its final performance, leads to an increase of the overall classification accuracy [11].

In a training process, the fuser needs to identify $W = \{W_1, W_2, \ldots, W_N\}$ where $W_l = \left[w^{(l)}(1), w^{(l)}(2), \ldots, w^{(l)}(M)\right]^T$ comprises the weights assigned to each classifier and each of the M classes.

The aim is to find a fuser which assures the lowest misclassification rate of $\bar{\Psi}$ for which we employ a neural network with a canonical learning approach [29] as illustrated in Fig. 1. One perceptron fuser is constructed for each of the classes under consideration, and may be trained with any standard procedure used in neural network learning. The input weights established during the learning process are then the weights assigned to each of the base classifiers.

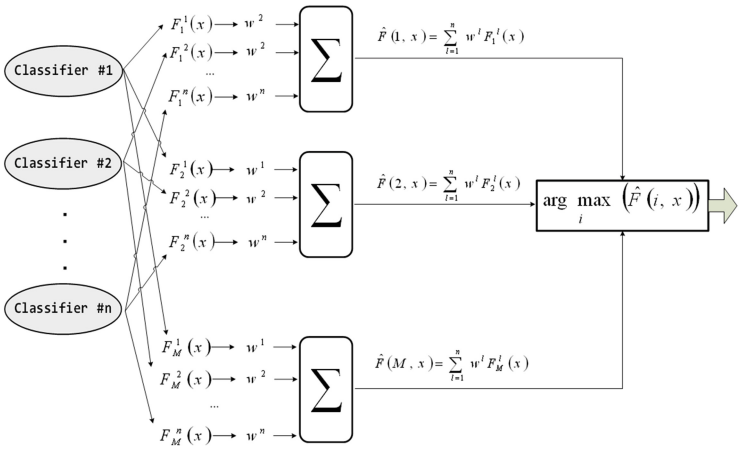

Fig. 1. Classifier fuser implemented as a one-layer neural network

5 Experimental Results

In our experiments, we use a dataset of 146 thermograms of which 29 cases have been confirmed as malignant whereas the other 117 cases were benign. The employed dataset is the same that was used in earlier work [18,24,31]. For all thermograms, the 38 features from Sect. 3 are extracted and serve as input for classification.

For our proposed classification approach, each subspace is designed so as to contain all objects from the minority class and an equal number of samples from the majority class, leading to a pool of 7–9 base classifiers (depending on the fold of CV). To observe the influence of the removal of redundant classifiers, we run our experiments with and without employing the diversity-based classifier selection. The pruned ensembles consist of 4–6 individual classifiers (again, depending on the fold of CV).

In order to put the obtained results into context, we also perform classification using several state-of-the-art ensembles dedicated to imbalanced classification, namely SMOTEBagging [28], SMOTEBoost [9], IIvotes [5] and EasyEnsemble [21], all with support vector machines (with a Gaussian RBF kernel and classifier tuning, as in our approach) as base classifiers. Furthermore, we run the experiments using several canonical classifiers, namely a single SVM [27], bagged SVM [7], boosted SVM [10], and Random Forest [8].

Classification results, based on 5×2 cross validation, are presented in Table 1 where, for each classifier and classifier ensemble, 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). We also perform a combined 5×2 CV F test of statistical significance (on sensitivity) [2], and report its results in Table 2.

Table 1. Classification results for all classifiers

Classifier	Sensitivity	Specificity	Accuracy
Single SVM [27]	8.34	86.32	71.23
Bagged SVM [7]	12.68	94.01	79.45
Boosted SVM [10]	19.58	98.00	85.61
Random Forest [8]	22.58	98.29	84.24
SMOTEBagging [28]	77.35	90.50	87.89
SMOTEBoost [9]	79.03	91.00	88.62
IIvotes [5]	79.56	91.89	89.44
EasyEnsemble [21]	80.02	91.00	88.22
Hybrid Ensemble (NN)	78.85	90.82	88.43
Hybrid Ensemble (SVM)	79.85	91.08	88.78
Hybrid Ensemble (NN) + DIV	80.74	90.52	88.56
Hybrid Ensemble (SVM) + DIV	81.96	90.80	89.03

From Table 1, we can see that canonical classification approaches are not able to cope well with the dataset due to the inherent class imbalance, and consequently provide rather poor sensitivity.

The implemented ensembles SMOTEBagging, SMOTEBoost, IIvotes and EasyEnsemble are all specifically designed to address class imbalance in the context of a multiple classifier system. SMOTEBagging and SMOTEBoost do this through oversampling approaches, while IIvotes is also based on the bagging combination idea, but integrates the SPIDER data preprocessing technique with the Ivotes approach. EasyEnsemble carries out a double ensemble learning procedure, combining bagging and boosting. From the results in Table 1, it is apparent that these methods lead to a significant boost in terms of sensitivity. Table 2 shows that all ensembles statistically outperform all single classifiers, while EasyEnsemble gives the best sensitivity.

Now looking at the results of our proposed hybrid ensemble approach, we can notice that it leads to a clear further improvement still. Using the employed subspacing method means that we do not need to create artificial objects or define cost matrices. In the former new artificial objects may be introduced on the basis of already created artificial samples, while for the latter a cost matrix needs to be defined which is often difficult and requires detailed domain knowledge.

Inspecting the differences for employing different base classifiers, it is clear that the SVM-based ensemble, which closely resembles our earlier approach introduced in [18], leads to (statistically) better classification performance. This may be caused by the fact that SVMs tend to work well on small datasets. Each subspace consists of a relatively small number of objects (all minority samples available in the fold and an equal number of majority ones), and consequently

Table 2. Results of statistical significance. A + signifies that the algorithm listed in this row statistically outperforms the algorithm listed in this column (based on sensitivity), a − indicates a statistically inferior performance.

	single SVM	bagged SVM	boosted SVM	Random Forest	SMOTEBagging	SMOTEBoost	IIvotes	EasyEnsemble	Hybrid Ensemble (NN)	Hybrid Ensemble (SVM)	Hybrid Ensemble (NN) + DIV	Hybrid Ensemble (SVM) + DIV
single SVM		−	−	−	−	−	−	−	−	−	−	−
bagged SVM	+		−	−	−	−	−	−	−	−	−	−
boosted SVM	+	+			−	−	−	−	−	−	−	−
Random Forest	+	+			−	−	−	−	−	−	−	−
SMOTEBagging	+	+	+	+		−	−		−		−	−
SMOTEBoost	+	+	+	+	+			−		−	−	−
IIvotes	+	+	+	+	+						−	−
EasyEnsemble	+	+	+	+		+			+		−	−
Hybrid Ensemble (NN)	+	+	+	+	+			−		−	−	−
Hybrid Ensemble (SVM)	+	+	+	+		+			+		−	−
Hybrid Ensemble (NN) + DIV	+	+	+	+	+	+	+	+	+	+		−
Hybrid Ensemble (SVM) + DIV	+	+	+	+	+	+	+	+	+	+	+	

NNs tend to be prone to overfitting in these subspaces, while SVMs are able to handle the dichotomisation process more effectively.

The proposed hybridisation with the fuzzy diversity measure introduced in this paper is shown to give our method a further edge. Comparing the ensembles with and without the classifier selection stage, it is obvious that the former achieve higher classification and sensitivity performance. The overall best results are achieved by an ensemble of support vector machines that gives the maximum diversity using our energy-based measure. This approach yields a sensitivity of 81.96 % which is shown to be statistically better than those of all other methods, while resulting in only a slight drop in terms of specificity, and confirms that our hybrid ensemble algorithm provides an excellent classification method.

The presented approach also clearly outperforms earlier approaches in the literature of breast thermogram analysis. In [24], the same features and dataset were employed together with a cost-sensitive fuzzy if-then rule based classifier optimised by a genetic algorithm and giving a sensitivity of 79.86 % with a specificity of 79.49 %. [31] also used the same data and features and reported a sensitivity and specificity of 79 %.

6 Conclusions

In this paper, we have presented an effective approach to analysing breast thermograms for cancer diagnosis. We extract a set of image features describing bilateral (a)symmetry between the two breast regions from the images, and use these as input to a pattern classification stage. Based on our earlier work, we create a classifier ensemble for classification and address class imbalance by training its base classifiers on balanced data subsets. Using support vector machines and neural networks as individual classifiers and a trained perceptron as classifier fuser, this is shown to provide a powerful decision making system as experimental results on a dataset of about 150 thermograms demonstrate. Crucially though, we additionally perform a classifier selection stage based on a fuzzy diversity measure to eliminate redundant classifiers and identify the best models, and confirm this to lead to even (statistically significant) better classification performance and to yield an ensemble that outperforms not only various canonical classifiers but also several ensemble classifiers designed to address class imbalance.

References

1. Acharya, U.R., Ng, E.Y.K., Tan, J.H., Sree, S.V.: Thermography based breast cancer detection using texture features and support vector machine. J. Med. Syst. **36**(3), 1503–1510 (2012)
2. Alpaydin, E.: Combined 5×2 CV F test for comparing supervised classification learning algorithms. Neural Comput. **11**(8), 1885–1892 (1999)
3. Anbar, N., Milescu, L., Naumov, A., Brown, C., Button, T., Carly, C., AlDulaimi, K.: Detection of cancerous breasts by dynamic area telethermometry. IEEE Eng. Med. Biol. Mag. **20**(5), 80–91 (2001)
4. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1995)
5. Błaszczyński, J., Deckert, M., Stefanowski, J., Wilk, S.: Integrating selective preprocessing of imbalanced data with Ivotes ensemble. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 148–157. Springer, Heidelberg (2010)
6. Boquete, L., Ortega, S., Miguel-Jimnez, J.M., Rodrguez-Ascariz, J.M., Blanco, R.: Automated detection of breast cancer in thermal infrared images, based on independent component analysis. J. Med. Syst. **36**(1), 103–111 (2012)
7. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
8. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
9. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: improving prediction of the minority class in boosting. In: 7th European Conference on Principles and Practice of Knowledge Discovery in Database, pp. 107–119 (2003)
10. Dong, Y., Han, K.: Boosting SVM classifiers by ensemble. In: 14th International World Wide Web Conference, pp. 1072–1073 (2005)
11. Duin, R.P.W.: The combining classifier: to train or not to train? In: 16th International Conference on Pattern Recognition, vol. 2, pp. 765–770 (2002)
12. Gautherie, M.: Thermobiological assessment of benign and maligant breast diseases. Am. J. Obstet. Gynecol. **147**(8), 861–869 (1983)

13. Head, J.F., Wang, F., Lipari, C.A., Elliott, R.L.: The important role of infrared imaging in breast cancer. IEEE Eng. Med. Biol. Mag. **19**, 52–57 (2000)
14. Jones, B.F.: A reappraisal of infrared thermal image analysis for medicine. IEEE Trans. Med. Imag. **17**(6), 1019–1027 (1998)
15. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: Kernlab an S4 package for kernel methods in R. J. Stat. Softw. **11**(9), 1–20 (2004)
16. Keyserlingk, J.R., Ahlgren, P.D., Yu, E., Belliveau, N., Yassa, M.: Functional infrared imaging of the breast. IEEE Eng. Med. Biol. Mag. **19**(3), 30–41 (2000)
17. Krawczyk, B.: Diversity in ensembles for one-class classification. In: Pechenizkiy, M., Wojciechowski, M. (eds.) New Trends in Databases and Information Systems. Advances in Intelligent Systems and Computing, vol. 185, pp. 119–129. Springer, Heidelberg (2012)
18. Krawczyk, B., Schaefer, G.: Evolutionary multiple classifier system based on space partitioning for breast thermogram analysis. In: 16th Online World Conference on Soft Computing in Industrial Applications (2011)
19. Krawczyk, B., Wozniak, M., Schaefer, G.: Improving minority class prediction using cost-sensitive ensembles. In: 16th Online World Conference on Soft Computing in Industrial Applications (2011)
20. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, Hoboken (2004)
21. Liu, X., Wu, J., Zhou, Z.: Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. B Cybern. **39**(2), 539–550 (2009)
22. Mookiaha, M.R.K., Acharyaa, U.R., Ng, E.Y.K.: Data mining technique for breast cancer detection in thermograms using hybrid feature extraction strategy. Quant. InfraRed Thermography J. **9**(2), 151–165 (2013)
23. Qi, H., Snyder, W.E., Head, J.F., Elliott, R.L.: Detecting breast cancer from infrared images by asymmetry analysis. In: 22nd IEEE International Conference on Engineering in Medicine and Biology (2000)
24. Schaefer, G., Zavisek, M., Nakashima, T.: Thermography based breast cancer analysis using statistical features and fuzzy classification. Pattern Recogn. **42**(6), 1133–1137 (2009)
25. Tan, T.Z., Quek, C., Ng, G.S., Ng, E.Y.K.: A novel cognitive interpretation of breast cancer thermography with complementary learning fuzzy neural memory structure. Expert Syst. Appl. **33**(3), 652–666 (2007)
26. Tang, X., Ding, H., Yuan, Y., Wang, Q.: Morphological measurement of localized temperature increase amplitudes in breast infrared thermograms and its clinical application. Biomed. Sig. Process. Control **3**, 312–318 (2008)
27. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
28. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: IEEE Symposium on Computational Intelligence and Data Mining, pp. 324–331 (2009)
29. Wozniak, M., Zmyslony, M.: Designing combining classifier with trained fuser - analytical and experimental evaluation. Neural Netw. World **20**(7), 925–934 (2010)
30. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. In: 20th International Conference on Machine Learning, pp. 856–863 (2003)
31. Zavisek, M., Drastich, A.: Thermogram classification in breast cancer detection. In: 3rd European Medical and Biological Engineering Conference, pp. 1727–1983 (2005)