Margin-based Refinement for Support-Vector-Machine Classification

Helene Dörksen and Volker Lohweg
inIT – Institute Industrial IT, Ostwestfalen-Lippe University of Applied Sciences, Lemgo, Germany
{*helene.doerksen, volker.lohweg*}*@hs-owl.de*

Keywords: Refinement of Classification, Robust Classification, Classification within Small/Incomplete Samples.

Abstract: In real-world scenarios it is not always possible to generate an appropriate number of measured objects for machine learning tasks. At the learning stage, for small/incomplete datasets it is nonetheless often possible to get high accuracies for several arbitrarily chosen classifiers. The fact is that many classifiers might perform accurately, but decision boundaries might be inadequate. In this situation, the decision supported by margin-like characteristics for the discrimination of classes might be taken into account. Accuracy as an exclusive measure is often not sufficient. To contribute to the solution of this problem, we present a margin-based approach originated from an existing refinement procedure. In our method, margin value is considered as optimisation criterion for the refinement of *SVM* models. The performance of the approach is evaluated on a real-world application dataset for *Motor Drive Diagnosis* coming from the field of intelligent autonomous systems in the context of Industry 4.0 paradigm as well as on several UCI Repository samples with different numbers of features and objects.

1 INTRODUCTION

Machine Learning (ML) in the context of robust classification becomes increasingly important for data and signal processing in modern complex application environments. As example, such environments might be networked Cyber-Physical-Systems (CPS) and Multi-sensor or Information Fusion Systems installed for machine analysis and diagnosis for Industry 4.0 (Niggemann and Lohweg, 2015), cognitive radio (Ahmad et al., 2010), mobile health (Yi et al., 2014), etc. In addition to the robust classification problem, in many real-world scenarios the following problem occurs: a sufficient amount of training data – depending on an adequate model – is not available, that is, classification tasks have to be operated even in the situation where it is not possible to run enough measurements to generate objects for an appropriate classifier training.

Under these constraints, established signal processing algorithms and systems are at their frontiers, if not even inapplicable in, e.g., scenarios with resource limitations, like embedded systems. The findings presented here utilize a recently published robust classifier optimisation methodology (Dörksen and Lohweg, 2014; Dörksen et al., 2014). In (Dörksen et al., 2014) an approach is presented where a classifier has to deliver trustful results under the opposite constraint, i.e. large amount of objects in the samples.

The topics regarding small/incomplete samples as well as optimisation in ML are of the certain interest and not new in the academic area (cf. (Chapelle et al., 2002a) and (Sra et al., 2012)). In the context of the contribution to the topics, we rely on *SVM*-based classifier concepts (Boser et al., 1992; Cortes and Vapnik, 1995; Vapnik, 1995) and present a new margin-based approach for classification. In our method, margin value is considered as optimisation criterion for the refinement of *SVM* models. To increase the margin of the separation, we employ a methodology called *ComRef* (Dörksen and Lohweg, 2014). *ComRef* is a combinatorial refinement method (optimisation) for classification tasks. From the time complexity point of view, in this frame proposed margin-based method requires two *SVM* computations and additional calculation related to the explained below *min/max rule*, which depends linearly on the number of features. Furthermore, in the experimental part of our paper, we show that for many samples margin-based refinement possesses higher generalisation ability as the initial *SVM* discrimination. Moreover, the results demonstrate, that the technique is especially suitable for small/incomplete data sets, i.e., such sets with a scarce number of objects for the description of classes.

As supplement to the discussed topic above, in the case a dataset is small/incomplete it is nonetheless often possible to get a high accuracy in the terms

Dörksen, H. and Lohweg, V.
Margin-based Refinement for Support-Vector-Machine Classification.
DOI: 10.5220/0006115502930300
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 293-300
ISBN: 978-989-758-222-6

of classification rates not only for one single classifier. For small samples many classifiers might perform accurately in terms of trained and even test data, however, the decision boundary might be inadequate. Especially, if real-world data processing is required and unknown data appear, within accuracy tests chosen classifier might fail. It is clear that in this case the discrimination supported by *margin-optimised characteristics* is more convenient than the decision accuracy alone.

We show our approach in experimental results for the scenario of small datasets. They are modelled for the dataset from the real-world industrial application *Motor Drive Diagnosis* (Bayer et al., 2013) and for UCI samples. For modelling of different grades of incompleteness in the data, several *K*-fold cross-validation tests (Alpaydin, 2010) are performed.

The paper is organised as follows. In Sec. 2 we present related work covering *multiple classification*, *margin* initiated classification and *dimensionality reduction*. The theoretical aspects of the approach will be described in Sec. 3. The experimental validation will be given in Sec. 4. Finally, the conclusions and future work will be discussed in Sec. 5.

2 RELATED WORK

As previously mentioned in the Introduction, our approach originates from (Dörksen and Lohweg, 2014). The shortcoming of original is its combinatorial structure, i.e. it is hardly applicable to the samples with large number of features. We overcome this problem by defining rules for refinement, which will be presented in Sec. 3.

Our method belongs to the class of *multiple classifiers*, such as, e.g. *boosting methods* (Freund and Schapire, 1996) or *neural networks* (Hagan et al., 1996) with more than one layer. The relation is the following: by the initial feature combination and down-streamed classification in the low-dimensional space, we are able to combine two or more classifiers. However, in general, the principle idea of our approach differs from the such of multiple learners.

Furthermore, state of art is composing multiple learners (Bag-of-classifiers, Bag-of-Feature concepts, Ensemble Learners) that complement each other to obtain higher classification accuracy. A survey on combining classifiers is given in (Kuncheva, 2004); some latest developments are found in (Zhou et al., 2013).

Due to the fact that *SVM* is a margin-based classifier, the main scope of the scientific developments originated in (Boser et al., 1992; Cortes and Vapnik, 1995; Vapnik, 1995) and derived from *SVM* foundations might be considered as related to our work. In this sense, some recent investigations which focus on advantages and applications of *SVM*, are presented, e.g., in the book (Ma and Guo, 2014).

State of art developments of margin-based techniques are such originated by *Multiple Kernel Learning* (Chapelle et al., 2002b) or, e.g., radius-margin-based methods presented in (Do et al., 2009). Further, there is a number of publications describing *Large-Margin-Nearest-Neighbour Classification* (cf. (Weinberger and Saul, 2009)).

Dimensionality reduction (DR) methods are related to our work. In general, *ComRef* method which we extend to the margin-based refinement, originates from DR. From the ML point of view, DR is often motivated by the increasing of the generalisation ability (Guyon et al., 2006). DR is based on the problem of selecting subsets of most useful, i.e., relevant and informative features, and ignoring the rest. Other motivations of the DR are, e.g., data reduction and data understanding. A profound survey regarding DR domain can be found in (Benner et al., 2005); nonlinear dimensionality reduction methods (Lee and Verleysen, 2007); some recently published studies are presented in (Pei et al., 2013).

In the variety of DR methods, the most closely related ones w.r.t. our approach are the *feature weighting* methods. A survey on the weighting methods is found in (Blum, A. L. and Langley, P., 1997).

3 APPROACH

In the framework of this contribution, we restrict our considerations to classifiers whose decision rules can be represented as weighted feature combinations. Due to the margin-based nature of the proposed method, we concentrate on the *SVM* model. However, in general, we expect our approach to be applicable to other classification models, e.g., *LDA* or *PCA* with a suitable definition of a *margin*.

The basic idea of the *SVM* model is to maximize the *margin*, which is the distance from the hyperplane to the objects closest to it on either side (Alpaydin, 2010; Cortes and Vapnik, 1995). Thus, the *margin* plays a crucial role in the design of *support-vectors-based* (SV) learning algorithms (Schölkopf and Smola, 2002).

The idea of our proposal is to increase the *margin* without increasing the complexity of the model (*VCdim* (Vapnik, 1995)). In that context, the complexity of the model is determined by the profile of the classification boundary with respect to the number of features or type of the kernel, i.e., higher number of features implies higher complexity, or e.g., quadratic kernel classifier possesses higher complexity as linear kernel classifier. In many cases, lower complexity implies higher robustness of classification, that can lead to better generalisation ability. From this point of view, our proposal improves classification rates staying robust.

A *Classification problem* of two classes T^+ and T^- with corresponding objects $\mathbf{x}^+ \in T^+$ and $\mathbf{x}^- \in T^-$ is considered. Objects are vectors $\mathbf{x} \in \mathbb{R}^d$ with $\mathbf{x} = (x_1, \cdots, x_d)$ in the d-dimensional feature space $X \subseteq \mathbb{R}^d$. We assume that a linear combination h of features is given:

$$h(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle = \sum_{i=1}^{d} a_i x_i. \qquad (1)$$

With some scalar $c \in \mathbb{R}$, the rule for the linear classification, w.r.t. Eq. (1) is the following:

$$\mathbf{x} \in T^+ \text{ if } h(\mathbf{x}) \geq c \quad \text{and} \quad \mathbf{x} \in T^- \text{ if } h(\mathbf{x}) < c. \quad (2)$$

Within our approach we do not distinguish between linear separable and non-separable cases. We concentrate only on the parameters provided by weighting vector $\mathbf{a}$. Let $h(\mathbf{x})$ be an *SVM* classifier. Thus, for the separable case in canonical form the so-named *margin* $\rho := 2/\|\mathbf{a}\|$ is maximized. Within labels $y = 1$ for all $\mathbf{x}^+$ and $y = -1$ for all $\mathbf{x}^-$, it is equivalent to the solution of the problem:

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{a}\|^2 \\ &\text{subject to } y_j\left(\langle \mathbf{a}, \mathbf{x}_j \rangle - c\right) \geq 1 \\ &\text{with } \mathbf{x}_j \in \{\mathbf{x}^+, \mathbf{x}^-\}, \forall j. \end{aligned}$$

For the non-separable case *slack variables* $\xi_j \geq 0$, $\forall j$ are defined. Slack variables store the deviation from the margin ρ in order to relax the constraints. A *soft margin* classifier for a non-separable case is the solution of the problem:

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{a}\|^2 + C\sum_j \xi_j \\ &\text{subject to } y_j\left(\langle \mathbf{a}, \mathbf{x}_j \rangle - c\right) \geq 1 - \xi_j \\ &\text{with } \mathbf{x}_j \in \{\mathbf{x}^+, \mathbf{x}^-\}, \xi_j \geq 0, \forall j, \end{aligned}$$

where the constant $C > 0$ determines the trade-off between margin maximization and training error minimization. Similar to separable case, for simplicity, the margin here is defined as $\rho := 2/\|\mathbf{a}\|$.

In the sense of classical *SVM* fundamentals, the functional $h(\mathbf{x})$ represents a hyperplane having the largest separation, or margin ρ, between two classes. The hyperplane has the property that the distance is maximized from the *nearest data point* on each side. Our refinement approach is based on the weakening of this property: the distance from the hyperplane to *some appropriate data* is maximized. This idea originates on the method from *Combinatorial Refinement*, called *ComRef*, approach (Dörksen and Lohweg, 2014), where it was shown that for many samples the refinement of the hyperplane is able to increase the generalisation ability of the classification. We build up our method on the well-known fact, that increasing of the generalisation ability might rely on the regularisation property of *SVM* hyperplane, i.e. on the property to find largest separation ρ.

Let us consider *SVM* classifier $h(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle$ with the initial margin ρ. In the frame of this paper, without loss of generality, we consider l_2-norm, i.e. :

$$\rho = \frac{2}{\|\mathbf{a}\|} = \frac{2}{\sqrt{a_1^2 + \cdots + a_d^2}}. \qquad (3)$$

Knowing $h(\mathbf{x})$, we are interested in the computation of a new one hyperplane $g(\mathbf{u}) = \langle \mathbf{b}, \mathbf{u} \rangle$, which will be called *refinement* of h. The margin ρ_{ref} of g has to be larger as ρ, i.e. $\rho_{ref} > \rho$. It is clear that, since $\mathbf{x}$ and $\mathbf{u}$ belong to dissimilar feature spaces, margins provided by h and g have to be comparable. We will discuss this topic later.

Within all assumptions and descriptions above, *fusion of summands* of h for some indices $I \subseteq \{1, \cdots, d\}$ is defined as:

$$u_I := \sum_{i \in I} a_i x_i.$$

For $I = \{1, \cdots, d\}$, the fusion of summands is h itself. Otherwise, h can be represented by I together with the fusion of summands for the indices $\bar{I} = \{1, \cdots, d\} \setminus I$, which are complementary to I. More generally, h can be represented by fusion of summands (for ease of use read: $u_{I_k} = u_k, k \in \mathbb{N}$):

$$h(\mathbf{x}) = \sum_{i \in I_1} a_i x_i + \cdots + \sum_{i \in I_j} a_i x_i = u_1 + \cdots + u_j, \quad (4)$$

where for $k = 1, \cdots, j$ holds $I_k \subset \{1, \cdots, d\}$ and all I_k are non-empty disjointed subsets of indices with the property that:

$$\bigcup_{k=1}^{j} I_k = \{1, \cdots, d\}.$$

For some parameters $b_1, \cdots, b_j \in \mathbb{R}$, *refinement of a linear classifier* resp. feature weighting of Eq. (1) and fusions of summands in Eq. (4) is defined as:

$$g(\mathbf{u}) = \langle \mathbf{b}, \mathbf{u} \rangle = \sum_{i=1}^{j} b_i u_i. \qquad (5)$$

From the definition above, it is clear that the refinement is performed in low-dimensional space $U \subseteq \mathbb{R}^j$

with $j < d$. With some scalar $\tilde{c} \in \mathbb{R}$, the rule for the linear classification resp. refinement Eq. (5) is now:

$$\mathbf{u} \in T^{+} \text{ if } g(\mathbf{u}) \geq \tilde{c} \quad \text{and} \quad \mathbf{u} \in T^{-} \text{ if } g(\mathbf{u}) < \tilde{c}. \quad (6)$$

In the framework of original *ComRef*, *SVM* (linear and quadratic) and *AdaBoost* (Freund and Schapire, 1996) classifiers for initial feature combination w.r.t. Eq. (1) were analysed. For many datasets it was shown, that fusion of summands, where the classification accuracy with respect to Eq. (6) is higher than in Eq. (2), might lead to the higher generalisation ability of Eq. (6) than of Eq. (2).

The main disadvantage of *ComRef* is its combinatorial nature in the term of selection of fusions of summands, i.e., for samples with large number of features the algorithm is hardly applicable. A technique for fast searching for a suitable fusion of summands is required. In the frame of the work here, we solve this problem by applying margin value as optimisation criterion. Thus, fusions of summands leading to larger margin in the refinement Eq. (5) will be considered.

Let $g(\mathbf{u})$ in Eq. (5) be an *SVM* classifier and be a refinement of h. Hence, the margin of g in the low-dimensional space is equal to $2/\|\mathbf{b}\|$. It is clear that for the construction of a margin-based optimisation criterion we are not able to compare both margins directly, since they are located in dissimilar spaces. To illustrate this effect, consider the hyperplane $\langle \mathbf{b}, \mathbf{u} \rangle - \tilde{c} = 0$ in the space $\mathbb{R}^{j}$. It corresponds to the representation in the initial space $\mathbb{R}^{d}$:

$$b_1 \sum_{i \in I_1} a_i x_i + \cdots + b_j \sum_{i \in I_j} a_i x_i - \tilde{c} = 0. \qquad (7)$$

Thus, under the assumption that $b_1 = 1, \ldots, b_j = 1$ and $c = \tilde{c}$, the hyperplanes in $\mathbb{R}^{d}$ and $\mathbb{R}^{j}$ are equivalent. However, the margin $\tilde{\rho} = 2/\sqrt{j}$ is, in general, not equal to $\rho = 2/\|\mathbf{a}\|$. Thus, based on the representation of Eq. (7), we define ρ_{ref} to be compared with initial margin $\rho = 2/\|\mathbf{a}\|$ for the optimisation criterion as follows:

$$\rho_{ref} = \frac{2}{\sqrt{b_1^2 \sum_{i \in I_1} a_i^2 + \cdots + b_j^2 \sum_{i \in I_j} a_i^2}}. \qquad (8)$$

Proposition. Let $h(\mathbf{x})$ in Eq. (1) be an *SVM* classifier with margin ρ. Assume, fusions of summands are given w.r.t Eq. (4). Further, let $g(\mathbf{u})$ in Eq. (5) be a refinement of h and be an *SVM* classifier with margin ρ_{ref}.

If in Eq. (8) for each $k = 1, \cdots, j$ holds $b_k^2 \leq 1$ then $\rho_{ref} \geq \rho$.

Proof. It can be easily seen, that for above assumptions the following is true:

$$\sqrt{a_1^2 + \cdots + a_d^2} \geq \sqrt{b_1^2 \sum_{i \in I_1} a_i^2 + \cdots + b_j^2 \sum_{i \in I_j} a_i^2}.$$

Thus, $\rho_{ref} \geq \rho$. □

Several techniques can be considered for the construction of fusion of summands, e.g. initiated by feature extraction or weighting methods. In the frame of our work, we discuss the simplest situation: *one single* fusion of *two summands*. It leads to the defined below *min/max rule* and, e.g. can be extended or integrated into refinement process iteratively. By *one single* fusion of *two summands*, the refinement is based on the representation of initial *SVM*:

$$h(\mathbf{x}) = a_1 x_1 + \cdots + \sum_{i \in I_k} a_i x_i + \cdots + a_d x_d, \qquad (9)$$

where I_k is a set of two indices from $\{1, \cdots, d\}$. It is clear, that the refinement occurs in $(d-1)$-dimensional space. Our explanation for paying attention to one single fusion of two summands is following: By the selection $\sum_{i \in I_k} a_i^2$ with only two summands we might expect that the refinement for a_i's, $i \notin I_k$, is marginal, i.e. $b_j \approx 1$, for $j \neq k$. In that case the margin value ρ_{ref} depends mainly on $\sum_{i \in I_k} a_i^2$ and b_k^2. It holds that, within optimisation of $\sum_{i \in I_k} a_i^2$ and b_k^2, the margin ρ_{ref} represented in Eq. (8) is more larger more smaller are $\sum_{i \in I_k} a_i^2$ and b_k^2. We deduce that $|a_i|$'s with small values might stronger enlarge the margin as such $|a_i|$'s with large values. On the other hand, since we are interested in margin optimisation, we can try to degrade the contribution of $|a_i|$'s with large values and in such way increase the margin.

Due to above discussions, the *min/max rules* (resp. for MIN or MAX later in the tables from Section 4) of *one single* fusion of *two summands* for margin-based refinement are following:

Min/Max Rules for Margin-based Refinement

i. Compute initial *SVM* hyperplane $h(\mathbf{x})$ as in Eq. (1).

ii. Find set I_k of two indices from $\{1, \cdots, d\}$ such that $\sum_{i \in I_k} a_i^2$ is minimal/maximal, i.e. choose from $|a_1|, \cdots, |a_d|$ two with mininimal/maximal values.

iii. Recalculate *SVM* refinement $g(\mathbf{u})$ for $\mathbf{u} = \big(a_1 x_1, \cdots, \sum_{i \in I_k} a_i x_i, \cdots, a_d x_d\big)$.

Table 1: Results of *K-fold cv paired* t *test* for *SVM* and proposed margin-based refinement approach (indicated as *Ref*MIN or *Ref*MAX w.r.t. *min/max rules*) for datasets M-I and M-II are listed. In rows, overall accuracies (in %), t_{K-1}-statistics as well as margin ρ_{ref} for benchmarking are given. Here, for original margin is valid $\rho = 2$ for all samples.

K-fold	*SVM*	*Ref*MIN	t_{K-1}MIN	ρ_{ref}MIN	*Ref*MAX	t_{K-1}MAX	ρ_{ref}MAX
M-I ($K = 30$)	92.26	93.88	7.69	2.80	93.76	6.93	2.68
M-I ($K = 100$)	88.77	90.37	8.85	2.67	90.04	6.95	2.54
M-II ($K = 30$)	96.00	97.08	15.25	2.84	97.04	15.00	2.76
M-II ($K = 100$)	94.39	95.45	10.24	2.83	95.37	9.51	2.70

Table 2: Results of *K-fold cv paired* t *test* ($K = 10$) for *SVM* and proposed margin-based refinement approach (here, only *Ref*MIN w.r.t.*min rule*) are listed for UCI datasets. The remaining table caption corresponds to such of Table 1. Within* marked samples are considered for further iterations of refinement presented in Table 3 below.

dataset	# features	# objects	*SVM*	*Ref*MIN	t_{K-1}MIN	ρ_{ref}MIN
CNAE (classes 6 vs. 7)*	299	240	79.35	86.52	2.89	1.89
CNAE (classes 7 vs. 9)	333	240	86.75	92.96	2.92	1.70
Ecoli*	6	220	95.22	96.47	2.52	2.17
Heart	13	270	74.89	77.44	3.91	2.49
Promoters	57	212	74.64	78.32	5.45	2.30
Seeds	7	140	91.66	93.49	5.81	2.54
Splice (classes *E* vs. *I*)*	60	1535	84.66	87.13	7.04	2.16
Splice (classes *I* vs. *N*)*	60	2423	81.72	84.54	6.25	2.09

4 EXPERIMENTAL RESULTS

4.1 Motor Drive Diagnosis

We show the results for a real-world industrial application *Motor Drive Diagnosis* (Bayer et al., 2013). The application relates to an intelligent, autonomic synchronous motor drive which is used in several applications in transport systems at airports, conveyor belts, etc.

We consider two classes—*intact* and *anomaly*—for the motor condition which is a conclusive result, taking into account the use of adequate features (Bator et al., 2012). The defined range of typical defects in drive train applications might be, e.g., *ball bearings*, *axle displacement* or *inclination of gear-wheels*. For our test, datasets called *Motor I* (M-I) and *Motor II* (M-II) with 52 features and total 5,318 objects in M-I resp. 72 features and 10,638 objects in M-II are analysed. The number of objects in M-II might appear large, however, for this type of applications it is still not complete. In general, it is almost impossible to collect complete and well-balanced samples for motor diagnosis. Due to the different environmental conditions (e.g. stable/unstable position of the motor, air temperature/humidity, running time, etc.), the real completeness/balance of the sample is assumed to be unknown. To show that our approach is able to contribute to the solution of the problem, we perform *K-fold cv paired* t *test* setting $K = 30$ and $K = 100$, i.e., one third (circa 3.3 %) resp. 1 % of information about the sample as available. We present the classification rates in terms of accuracies (in %) and t_{K-1}-*statistics* (Alpaydin, 2010) for comparing two classification algorithms. The larger is the value of the t_{K-1}-statistic, the more likely one algorithm is performing better/worse as the other. For t_9 it holds: If $|t_9| > 2.26$, then we reject the hypothesis that the algorithms have the same error rate with 97.5% confidence. For $|t_{29}|$ it is 2.05, hence, for $|t_{99}|$ it is lower than 2.05.

Without loss of generality, *SVM* separating hyperplane is computed using *Sequential Minimal Optimization* (Schölkopf and Smola, 2002). Furthermore, all data sets are standardised to have margin $\rho = 2$ in initial *SVM*; features are standardised to have equal standard deviations $\sigma^2 = 1$; positions of objects in each set are randomised before selecting training and testing subsets. The results, presented in Table 1, illustrate the performance. We remark that the accuracies in the learning stage were close to 100% for both initial *SVM* and refinement. Differently to accuracies, the margin values of refinements were larger as such of original *SVM*. Due to increasing of the margin, the overall accuracy improvement of the refinement is between $1\% - 2\%$, that is a good achievement for industrial environments taking into account huge networked Cyber-Physical-Systems and the financial profitability from this kind of optimisation.

Table 3: Results of *K-fold cv paired* t *test* ($K = 10$) for two successive iterations of *min rule*.

dataset	*SVM*	Iter.*Ref*MIN	t_{K-1}MIN	ρ_{ref}MIN
CNAE (classes 6 vs. 7)	79.35	90.04	4.64	1.95
Ecoli	95.22	97.23	3.49	2.20
Splice (classes *E* vs. *I*)	84.66	88.18	10.47	2.27
Splice (classes *I* vs. *N*)	81.72	86.23	6.34	2.14

Table 4: *F-measures* based on true positives and false negatives of each class for *K-fold cv paired* t *test* ($K = 10$).

dataset	# features	# objects	balance	$F_{SVM}(T^+ / T^-)$	$F_{RefMIN}(T^+ / T^-)$
Fertility	9	100	7.3	0.12 / 0.82	0.16 / 0.83
Hepatitis	16	80	5.2	0.32 / 0.84	0.35 / 0.85
Heart	13	150	4.0	0.47 / 0.84	0.51 / 0.84
Mammography	5	628	4.6	0.67 / 0.91	0.69 / 0.92
Pima	8	331	4.3	0.45 / 0.83	0.49 / 0.84
Promoters	57	120	7.6	0.04 / 0.93	0.14 / 0.93
Vertebral	6	244	6.2	0.40 / 0.86	0.44 / 0.86

4.2 UCI Repository Test Samples

In this section we present the experimental results of our proposed margin-based learning approach based on several UCI Repository samples (*http://mlr.cs.umass.edu/ml/datasets.html*). Following data sets are considered: *CNAE* (class labels 6 vs. 7, and 7 vs. 9), *Ecoli, Heart, Promoters, Seeds* and *Slice* (class labels *E* vs. *I*, and *I* vs. *N*). To show the potentiality of the method, we present results for samples with different numbers of features and objects, where margin-based refinement increases the generalisation ability of the classification. Since often it is not known if the sample is complete or not, there might be data where refinement is not working. We perform *K-fold cv paired* t *test* for generalisation (Alpaydin, 2010) with $K = 10$, i.e., in each fold 10% of the objects are considered for the training and the rest for the validation. We demonstrate only results for *min rules*, for many samples *max rule* is leading to improvement, however, as our tests show for considered samples, *min rule* is slightly stronger as *max rule*. We remark that also here for many considered samples the accuracies in the learning stage were close to 100% for both initial *SVM* and refinement. Table 2 represents the results. Results show that refinement rules lead often to the increasing of the margin. Due to increasing of the margin, the accuracy is increasing in the refinement as well.

In the sample *CNAE*, margin slightly decreases, however, refinement leads to higher accuracy.

In addition, we show some results for two successive iterations of *min rule* in Table 3.

Finally, we test our approach on several samples which are not well-balanced, meaning for our examples that numbers of objects in the classes are exceeding four times apart. Accuracy as an exclusive measure is here not sufficient as well. Scores based on true positives and false negatives of each class is suitable for such problems. Following UCI samples are considered: *Fertility, Hepatitis, Heart, Mammography, Pima, Promoters* and *Vertebral*. The samples *Fertility* and *Hepatitis* are not well-balanced in original UCI submission. In the remaining samples, we scale them artificially by removing objects from one of the classes, such that they become non-balanced. We define balance of the sample as number of objects of one class divided by the number of the objects of the other. *F-measures* are calculated as scores based on true positives and false negatives for each class, where:

$$F(class) = \frac{2\,TP(class)}{2\,TP(class) + FP(class) + FN(class)},$$

and *TP, FP, FN* are resp. true positives, false positives and false negatives. Table 4 represents the results.

5 CONCLUSIONS

In this paper we presented a fast approach for margin-based refinement of *SVM* classifier. We defined *min/max rules* for the refinement procedure and illustrated the performance on several samples with different numbers of features and objects. In our future work we investigate characteristics of fusions for being suitable for that kind of refinement. We examine additional rules for fast refinement. Furthermore, combinations of rules will be studied. We will perform the theoretical and empirical analysis of the refinement methodology in the context of feature selection and weighting.

In addition, comparison with other existing weighted *SVM* models will be evaluated. For efficiency reasons, possibilities for incorporation of refinement recalculations into one single step will be studied.

ACKNOWLEDGEMENTS

We would like to thank Martyna Bator of Institute Industrial IT for providing helpful support regarding Motor Drive Diagnosis dataset. This work was partly funded by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster ”Intelligent Technical Systems Ost-Westfalen Lippe” (its OWL).

REFERENCES

Ahmad, K., Meier, U., and Kwasnicka, H. (2010). Fuzzy logic based signal classification with cognitive radios for standard wireless technologies. In *Cognitive Radio Oriented Wireless Networks Communications (CROWNCOM), 2010 Proceedings of the Fifth International Conference on*, pages 1–5.

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, Cambridge, 2 edition.

Bator, M., Dicks, A., Mönks, U., and Lohweg, V. (2012). Feature Extraction and Reduction Applied to Sensorless Drive Diagnosis. In Hoffmann, F. and Hüllermeier, E., editors, *22. Workshop Computational Intelligence*, volume 45 of *Schriftenreihe des Instituts für Angewandte Informatik - Automatisierungstechnik am Karlsruher Institut für Technologie*, pages 163–178. KIT Scientific Publishing.

Bayer, C., Bator, M., Mönks, U., Dicks, A., Enge-Rosenblatt, O., and Lohweg, V. (2013). Sensorless Drive Diagnosis Using Automated Feature Extraction, Significance Ranking and Reduction. In Seatzu, C. and Zurawski, R., editors, *ETFA 2013*, pages 1–4.

Benner, P., Mehrmann, V., and Sorensen, D. C. D. C., editors (2005). *Dimension reduction of large-scale systems : proceedings of a workshop held in Oberwolfach, Germany, October 19-25, 2003*, Lecture notes in computational science and engineering. Springer.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT ’92, pages 144–152, New York, NY, USA. ACM.

Chapelle, O., Vapnik, V., and Bengio, Y. (2002a). Model selection for small sample regression. *Machine Learning*, 48(1-3):9–23.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002b). Choosing multiple parameters for support vector machines. *Mach. Learn.*, 46(1-3).

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Do, H., Kalousis, A., Woznica, A., and Hilario, M. (2009). Margin and radius based multiple kernel learning. In Buntine, W., Grobelnik, M., Mladenić, D., and Shawe-Taylor, J., editors, *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, Proceedings, Part I*. Springer, Berlin, Heidelberg.

Dörksen, H. and Lohweg, V. (2014). Combinatorial refinement of feature weighting for linear classification. In *Emerging Technology and Factory Automation (ETFA), 2014 IEEE*, pages 1–7.

Dörksen, H., Mönks, U., and Lohweg, V. (2014). Fast classification in industrial Big Data environments. In *Emerging Technology and Factory Automation (ETFA), 2014 IEEE*, pages 1–7.

Freund, Y. and Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. *Thirteenth International Conference on Machine Learning*.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2006). *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc, Secaucus and NJ and USA.

Hagan, M. T., Demuth, H. B., and Beale, M. (1996). *Neural Network Design*. PWS Publishing Co., Boston, MA, USA.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 1st edition.

Ma, Y. and Guo, G. (2014). *Support vector machines applications*. Springer.

Niggemann, O. and Lohweg, V. (2015). On the diagnosis of cyber-physical production systems. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 4119–4126.

Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G., editors (2013). *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I-III*, Lecture Notes in Computer Science. Springer.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, Cambridge and Mass.

Sra, S., Nowozin, S., and Wright, S. J. (2012). *Optimization for Machine Learning*. Neural information processing series. MIT Press.

Vapnik, V. N. (1995). *Statistical Learning Theory*. Wiley, London, 1 edition.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244.

Yi, W.-J., Sarkar, O., Mathavan, S., and Saniie, J. (2014). Wearable sensor data fusion for remote health assessment and fall detection. In *Electro/Information Technology (EIT), 2014 IEEE International Conference on*, pages 303–307.

Zhou, Z. H., Roli, F., and Kittler, J. (2013). *Multiple Classifier Systems: 11th International Workshop, MCS 2013, Nanjing, China, May 15-17, 2013. Proceedings*. Lecture Notes in Computer Science, vol. 7872. Springer London, Limited.