PROCEEDINGS

53. IWK

Internationales Wissenschaftliches Kolloquium
International Scientific Colloquium

Faculty of
Mechanical Engineering

PROSPECTS IN MECHANICAL ENGINEERING

8 - 12 September 2008

www.tu-ilmenau.de

TECHNISCHE UNIVERSITÄT
ILMENAU

Home / Index:
http://www.db-thueringen.de/servlets/DocumentServlet?id=17534

Published by
Impressum

Publisher Herausgeber	Der Rektor der Technischen Universität Ilmenau Univ.-Prof. Dr. rer. nat. habil. Dr. h. c. Prof. h. c. Peter Scharff
Editor Redaktion	Referat Marketing und Studentische Angelegenheiten Andrea Schneider Fakultät für Maschinenbau Univ.-Prof. Dr.-Ing. habil. Peter Kurz, Univ.-Prof. Dr.-Ing. habil. Rainer Grünwald, Univ.-Prof. Dr.-Ing. habil. Prof. h. c. Dr. h. c. mult. Gerd Jäger, Dr.-Ing Beate Schlütter, Dipl.-Ing. Silke Stauche
Editorial Deadline Redaktionsschluss	17. August 2008
Publishing House Verlag	Verlag ISLE, Betriebsstätte des ISLE e.V. Werner-von-Siemens-Str. 16, 98693 Ilmenau

CD-ROM-Version:

Implementation Realisierung	Technische Universität Ilmenau Christian Weigel, Helge Drumm
Production Herstellung	CDA Datenträger Albrechts GmbH, 98529 Suhl/Albrechts

ISBN: 978-3-938843-40-6 (CD-ROM-Version)

Online-Version:

Implementation Realisierung	Universitätsbibliothek Ilmenau ilmedia Postfach 10 05 65 98684 Ilmenau

Home / Index:
http://www.db-thueringen.de/servlets/DocumentServlet?id=17534

53rd Internationales Wissenschaftliches Kolloquium
Technische Universität Ilmenau
08 – 12 September 2008

K. Anding / D. Garten

Comparison of Different Classification Algorithms at the Application of Automatical Quality Assurance of Grain

ABSTRACT

Wheat is one of the most widely grown cereal crops in the world and a staple food in many countries. So the quality assurance of wheat plays a leading role in food-manufacturing, particularly with regard to the wheat production for human nutrition. The negative effects of some impurities (fungus-covered grain, toxic foreign seed, ergot) on human and animals are well known. The analysis of elements in a grain sample is called “Besatz analysis of cereal”. The standard of analysis is described in the ICC standard methods from the International Association for Cereal Science and Technology [1], [2]. The state of the art for Besatz analysis of cereal is the manual inspection from a grain sample by laboratory assistants or leading millers. This expensive, time-consuming and error-prone procedure should be automated by machine learning. An automated object recognition routine for the Besatz analysis of grain is the task to be solved. During this study different machine learning algorithms were tested on this complex recognition problem.

INTRODUCTION

In our studies an image of every single constituent of a wheat sample was acquired by a colour line scan camera. Samples of 14 object classes from different cereal varieties and different grain impurities of wheat are used: flawless wheat, broken wheat, wheat damaged by pests, sprouted wheat, shrivelled wheat, durum, oat, rape, rye, sunflower seeds, weed seeds, husks, stones and contaminations pre-classified by a human expert (see Figure 1).

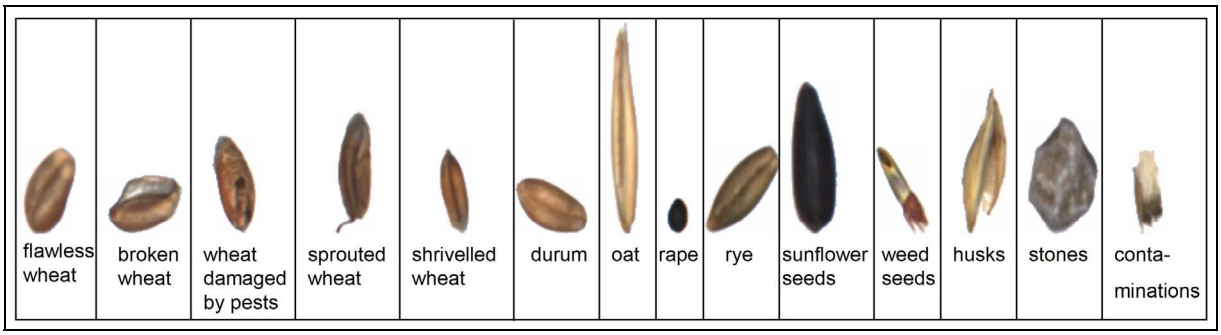

Figure 1: dataset samples for different object classes

After the image acquisition, the images were saved in bitmap-format. Images with significant shadows, artefacts and objects blurred by rotation were deleted. Afterwards all objects were randomly divided into training and test dataset. We used nearly 1000 objects per class for training and 500 for testing. For the classes “weed seeds” and “contaminations” there were only half of the number of objects available.

After segmentation and transformation from the RGB to HSI colour space a feature vector for every object has been calculated. We used 32 colour and texture features like the mean value per channel and features calculated from the co-occurrence matrix like energy, homogeneity and contrast per each HSI-channel. 92 scale and rotation-invariant shape features like modified Fourier descriptors were also calculated.

After feature extraction different classification algorithms were tested to determine their generalization ability. Experiments in feature selection were also conducted. We mainly used the machine learning toolbox “Weka 3.5” [3] and the commercial machine vision library Halcon 8.0. The Support-Vector-Machines (SVM) reached recognition rates at about 91% and about 99% for certain classes for this complex recognition problem. Other methods were much below.

RECOGNITION RATES OF DIFFERENT CLASSIFICATION ALGORITHMS

Choice of the optimal classification algorithm is an important task of research into classification of grain samples. So some classification algorithms were tested for a given dataset. The results, total recognition rates and recognition rates per each object class are shown in Figure 2, Table 1 and Table 2.

In our experiments Naïve Bayes reached very low recognition rates. This method treats all features as independent. Our experiments in feature selection indicated that there are significant correlations between many features. As a result of that Naïve Bayes is not a

suitable classifier. Experiments with algorithms from the LVQ family of algorithms were also not satisfactory.

The only method which reached nearly the recognition rates of the SVM is Random Forest which is an extension of the decision tree. Random Forest generates a very complex model. There is no advantage in classification time in comparison with the SVM but a lower recognition rate. As a result of that Random Forest was not further investigated.

The SVMs reached the highest recognition rates. This can be traced back to the special properties of the SVM. Later we will shortly describe the main features of the SVM-method.

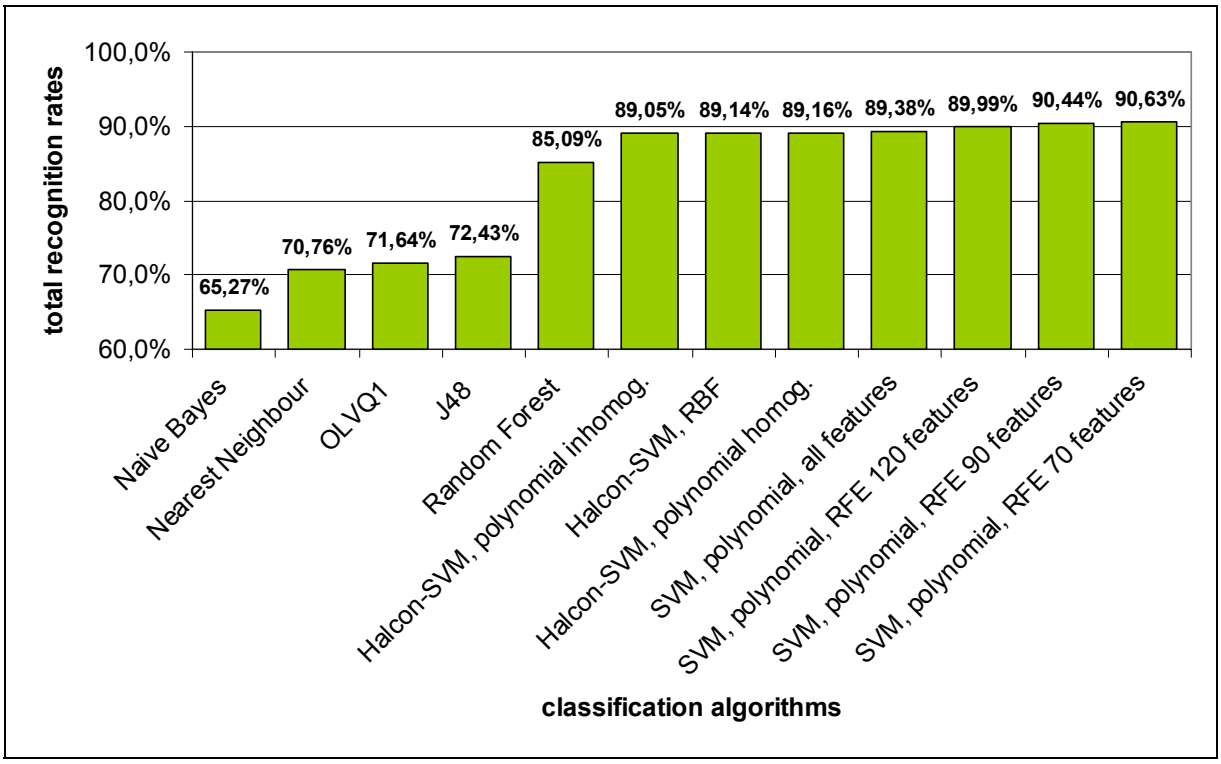

Figure 2: recognition rates with different classification algorithms

The wheat sub-classes are mostly difficult to separate because of their phenotypic similarity (see Table 1 and Figure 3). The most discriminatory features for the separation between the sub-classes of wheat (common wheat and hard wheat) are texture features. But because of the low resolution of our images, texture features are not useful. The objects are only imaged from one side. If a certain damage of a wheat kernel is only visible from one side and this side is turned away from the camera, we will get a false classification. This is another cause for bad recognition rates for the sub-classes of

wheat. Figure 2 shows the ability of different classification algorithms for discriminating between the 14 classes of sample dataset.

Model	recognition rate per class in %					
	common wheat					hard-wheat
	flawless wheat	broken wheat	wheat damaged by pests	sprouted wheat	shrivelled wheat	Durum
Naive Bayes	68,80	43,95	30,00	24,77	76,20	65,00
J48	57,60	56,25	42,40	45,18	71,80	77,20
Random Forest	75,80	73,19	71,60	68,81	85,60	91,00
Nearest neighbour	52,20	47,58	41,20	52,06	61,60	73,40
OLVQ1	53,80	56,05	40,60	50,20	57,40	76,60
SVM, poly. all features	79,80	85,48	77,40	81,88	86,20	95,80
SVM, poly. RFE 70 features	85,80	83,67	80,80	83,94	88,20	96,80
SVM, poly. RFE 90 features	83,40	85,89	80,60	85,32	87,40	96,40
SVM,poly. RFE 120 features	83,00	86,90	78,80	83,72	84,00	95,80
Halcon-SVM, RBF	82,80	81,80	78,80	79,13	85,40	94,60
Halcon-SVM, poly, homog.	79,00	83,00	79,00	77,75	86,00	95,60
Halcon-SVM, poly, inhomog.	80,40	81,80	78,80	78,67	84,40	95,40
Mean	73,53	72,13	65,00	67,62	79,52	87,80
standard deviation	12,31	16,33	19,90	19,74	10,53	11,37

Table 1: recognition rates of hard wheat and common wheat

Model	recognition rate per class in %							
	foreign grains				foreign seeds	impurities		
	oat	rape	rye	sun-flower seeds	weed seeds	husks	stones	conta-minations
Naive Bayes	78,80	96,65	66,60	94,00	69,23	65,80	77,6	44,71
J48	87,60	98,53	76,60	95,60	76,92	76,00	85	62,75
Random Forest	95,40	99,16	89,20	96,60	84,13	87,40	92,4	72,94
Nearest neighbour	87,40	96,86	76,80	96,00	81,25	78,60	77,6	55,69
OLVQ1	88,00	96,65	79,00	95,80	81,25	79,20	88,00	57,25
SVM, poly. all features	95,80	97,27	95,60	97,20	90,87	91,60	94	76,08
SVM, poly. RFE 70 features	96,00	98,74	94,80	98,80	91,83	91,80	93,6	78,04
SVM, poly. RFE 90 features	96,60	99,16	93,60	98,60	91,83	90,20	93,6	78,04
SVM,poly. RFE 120 features	96,40	99,16	93,80	98,20	92,31	90,40	94,4	78,04
Halcon-SVM, RBF	94,00	97,40	94,00	97,40	90,87	91,20	95,00	82,12
Halcon-SVM, poly, homog.	95,00	98,60	94,80	97,40	91,83	91,20	94,80	79,56
Halcon-SVM, poly, inhomog.	95,00	98,80	94,00	98,20	92,31	91,60	94,20	79,20
Mean	92,17	98,08	87,40	96,98	86,22	85,42	90,02	70,37
standard deviation	5,52	1,03	9,90	1,43	7,63	8,50	6,53	12,14

Table 2: recognition rates of non-wheat-elements

THE THEORY BEHIND SVM

The support vector machine (SVM) was introduced by [4] and is mentioned one of the

most powerful classifiers today. It is derived from statistical learning theory [5]. In our studies the SVM gave the highest recognition rates. The algorithm is motivated by structural risk minimization which says that not only the training error but also the complexity of the model influences its generalization ability. The SVM was designed to solve binary classification problems and can be modified for regression. There are different strategies to solve multi-class problems. We used the “one-versus-one” method. During the training process an optimal hyperplane is constructed. Optimal means that it leaves a maximal margin between the hyperplane and the closest training point on both sides. The hyperplane is defined by its coefficients α_i and the support vectors, which are determinated during the training process. Only these training vectors which are close to the boarder between the two classes become support vectors and hence have an influence on the resulting model. The decision function [8] to assign a class label {-1,1} to the feature vector x is given by:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{m} \alpha_i y_i x^t x_i + b) \tag{1}$$

with x_i as the i-th support vector and y_i as its class label {-1,1}, and b a predefined parameter. In the form mentioned above, the SVM can only give a linear boarder.

For constructing a non-linear SVM the so called “kernel trick” is used. The dot product $x^t x'$ in equation (1) can be replaced by a kernel function $k(x, x')$. This results in the following decision function [8]:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{m} \alpha_i y_i k(x, x_i) + b) \tag{2}$$

The kernel function $k(x, x_i)$ defines a scalar product in a higher dimensional space. In our studies we used the polynomial kernel [8]:

$$k(x, x') = [(x^T x') + a]^d \text{ with } a = \{0,1\} \tag{3}$$

and the Radial Basis Function kernel (RBF) [8]:

$$k(x, x') = e^{-\gamma \left\| x - x' \right\|^2} \tag{4}$$

The selection of the kernel function, its parameters and the complexity Parameter C which controls the balance between the training error and the margin are highly influencing the generalization ability. Choosing γ and C too big can result in over-fitting caused by a too complex model. And choosing it too small can give a too simple model. We found out that the selection of the degree d of the polynomial kernel as a small integer is easier and gives nearly the same performance like the selection of an

optimal γ for the RBF-kernel as a real number ranging from 0 to 1,0. For the selection of γ we need efficient optimization algorithms and a great amount of time because time for training a SVM model grows quadratic with the size of the training set.

ANALYSIS OF THE CONFUSION MATRIX

Figure 3 shows the confusion matrix calculated from the classification of the independent test set with the model Halcon-SVM with a homogeneously polynomial kernel function.

Halcon-SVM with homogeneously polynomial kernel function
dataset with 14 object classes
confusion matrix

classified as \ real class	flawless wheat (K1)	durum (K2)	oat (K3)	rye (K4)	stones (K5)	sprouted wheat (K6)	broken wheat (K7)	shrivelled wheat (K8)	wheat damaged by pests (K9)	rape (K10)	sun-flower seeds (K11)	husks (K12)	weed seeds (K13)	conta-minations (K14)
K1	**395**	8	0	0	0	**39**	**4**	**4**	**36**	0	0	0	0	0
K2	5	**478**	2	3	0	3	1	2	1	0	1	0	0	1
K3	0	0	**475**	0	0	0	0	0	0	0	1	11	9	2
K4	1	5	11	**474**	3	0	0	5	0	0	4	0	0	1
K5	0	0	0	0	**474**	2	2	0	0	0	1	2	0	2
K6	**33**	5	0	8	2	**339**	**11**	**4**	**25**	0	1	0	0	0
K7	**12**	0	2	1	3	**6**	**415**	**24**	**15**	0	0	4	0	4
K8	**4**	4	4	9	2	**12**	**36**	**430**	**28**	0	0	1	0	1
K9	**50**	0	0	2	2	**33**	**28**	**30**	**395**	0	1	0	0	0
K10	0	0	0	0	0	0	0	0	0	**493**	1	0	0	3
K11	0	0	0	1	2	0	0	0	0	0	**489**	0	0	0
K12	0	0	5	1	12	0	2	0	0	0	0	**456**	7	37
K13	0	0	0	1	0	0	0	1	0	3	0	3	**191**	5
K14	0	0	1	0	0	2	1	0	0	4	1	23	1	**218**

Figure 3: confusion matrix

Figure 3 shows wrong classifications above and below the main diagonal of the confusion matrix and correct classifications of the main diagonal.

The wheat sub-classes flawless wheat, wheat damaged by pests, sprouted wheat, broken wheat and shrivelled wheat are mostly difficult to separate from each other because of their phenotypic similarity (see Figure 3). These classes have almost identical colour and similar shape parameters. Therefore new feature algorithms have to be develloped, which were adapted for the specific recognition problem.

FEATURE SELECTION

At the given problem different feature selection methods respectively different data sets give different feature subsets. This indicates that many features are correlated. In our studies we used the support vector machine in conjunction with different feature selection strategies. The elimination of low rated features could not significantly increase the recognition rates. This indicates that the support vector machine does not suffer from the “Curse of Dimensionality” like simple local classifiers, for example Nearest-Neighbour, often do. In the future we will use ensemble methods to get a stable feature set and we will concentrate on the development of new features suitable for grain classification instead of eliminating useless ones. Because of the much higher recognition rates of the SVM in comparison with other learning algorithms, a combination of feature selection with other algorithms than SVM was not used.

We used ReliefF [6], Ambiguity [7] and SVM-RFE (Recursive feature elimination) [8] implemented in Weka [3]. Only with SVM-RFE, which is an embedded feature selection method for the SVM, a little increase of the recognition rates could be observed. Further investigation needs to be done to determine if it is a real advancement in generalization ability or over-fitting.

Figure 4 shows the slightly different feature rankings calculated with ReliefF and Ambiguity-Measure for our training dataset.

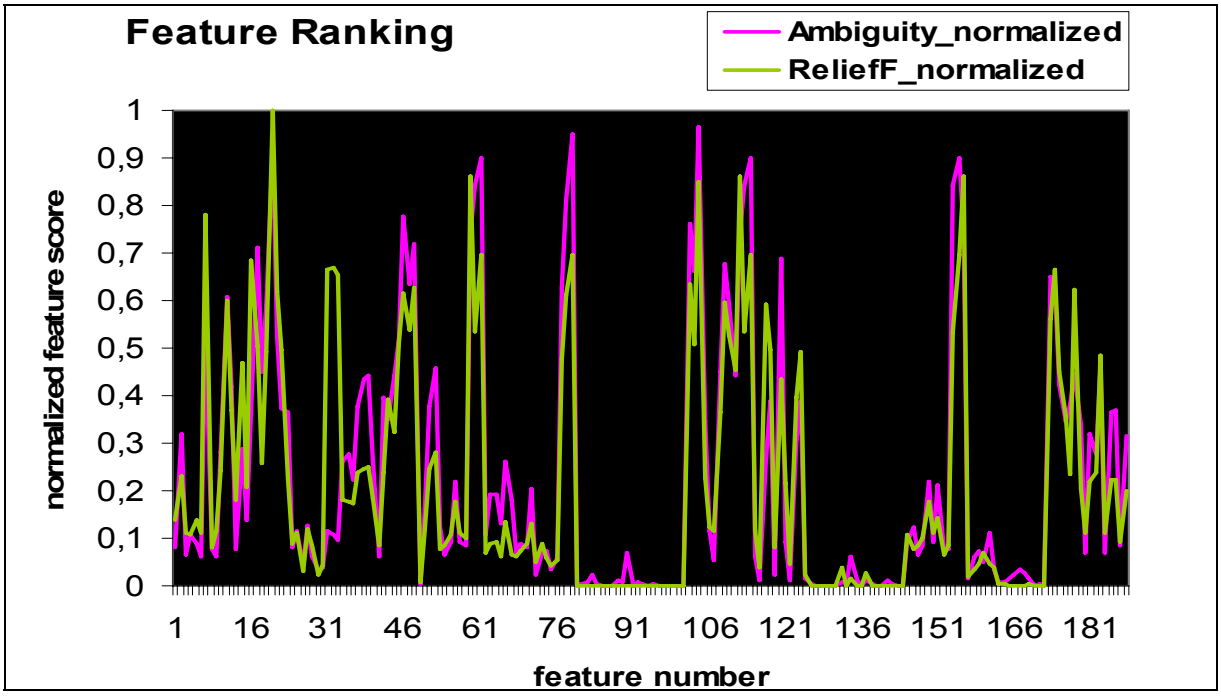

Figure 4: feature ranking calculated by ReliefF and Ambiguity-Measure

CONCLUSIONS

In this approach a part of relevant cereals and impurities, which can be contained in a grain sample, were used for testing different classification algorithms. The results especially the recognition rates have shown, that recognition of classes from natural products is an extensive and sophisticated challenge. Our approaches demonstrated SVM as the best classification algorithms for this recognition problem. The best SVM achieved a total recognition rate of about 90 % for this dataset. In future challenges we have to find new adapted feature algorithms and optimize the dataset. This means that we have to extend the dataset about other object classes and have to find the characteristic object features of each class.

ACKNOWLEDGEMENTS

A project funded by the Federal Ministry for Economic Affairs and Technology under the promotional reference 16INO496 forms the basis of this paper. The responsibility for the content of this paper lies with the author.

References:

[1] ICC No. 102/1: Determination of Besatz of Wheat. International Association for Cereal Science and Technology, 1972
[2] ICC No. 103/1: Determination of Besatz of Rye. International Association for Cereal Science and Technology, 1972
[3] I. Witten, E. Frank: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 2000
[4] B. Boser, I. Guyon, and V. Vapnik: A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning, 1992
[5] V. Vapnic: The Nature of Statistical Learning Theory. Springer, New York, 1995
[6] I. Kononenko: Estimating attributes: analysis and extensions of Relief. In: L. De Raedt and F. Bergadano (eds.): Machine Learning: ECML-94. pp. 171–182, Springer Verlag
[7] W. Abmayr: Einführung in die digitale Bildverarbeitung. B.G. Teubner Verlag Stuttgart, 1994
[8] I. Guyon et al.: Feature Extraction, Foundations and Applications. Physica-Verlag, Springer, 2006

Authors:

Dipl.-Ing. Katharina Anding
Dipl.-Inf. (FH) Daniel Garten
Technische Universität Ilmenau, Quality Assurance Department
P.O. Box 100 565, 98684 Ilmenau (Germany)
Phone: +49 3677 693964
Fax: +49 3677 693823
E-mail: katharina.anding@tu-ilmenau.de