Research Paper

Agricultural Science

Automatic Detection of Seed Quality and Varieties by Classification Methods

Selahaddin Batuhan AKBEN

Bahce Vocational School, Osmaniye, Turkey

ABSTRACT

In this study, automatic seed classification method is proposed according to the seed characteristics.Thus varieties and qualities of seeds will be able to identify without requiring analysis difficult to apply.The most important difference of this study from similar is proposition of method having easy to use in real life. To achieve this aim, different classification methods have been tested and evaluated by the experimental results. According to the experimental results by utilizing the geometric properties of the seeds, the most applicable classification method has been proposed for use in real-life.

KEYWORDS : Seed Classification, Machine Learning

Introduction

Seed analysis and classification are performed by agronomists and the other related experts by visually or chemical inspection of each sample (Luo et. al., 1999). It provides important knowledge to identification of impurities although time requiring task. Therefore automatically performing of it will be more convenient.Recently, very few researchers have proposed some automatic classification methods to use in seeds (Charytanowicz et. al., 2010; Zhao-yan et. Al., 2005; Granitto, et. al., 2002; Granitto, et. al., 2005). Some of these researches are related to clustering (Unsupervised learning) and others are related to classification methods (Supervised learning). Clustering methods can only use to find number of types of seeds or to find number of quality of seeds. Because clustering methods cannot be used for determination of name of findings.Furthermore in previous studies related to classification methods some methods are not mentioned.So the availability of classificationmethods in real life was not considered in previous studies. To eliminate this obscurity, most popular (well-known) classification methods should be evaluated by comparing in terms of performance and applicability.

Therefore the aim of this study is to compare classification methods in terms of performance and applicability for real life applications in seeds classification. In this study firstly the most well-known classification methods were applied to seeds data set. Then methods were compared with each other in terms of accuracy rate. Finally, high performance methods were evaluated in terms of applicability to real life and the most applicable method has been proposed.

Materials and Methods

Materials

For the ease of real life application the geometric characteristics of the seeds were used in this study. Because the botanical and chemical measurements requires stringent analysis and they are difficult to obtain these types of data. Therefore seed data set from UCI Machine Learning Repository was used in this study (Frank, A. et. al., 2014).Seed data set contains three different varieties of wheat kernels named as Kama, Rosa and Canadian.Parameters of wheat kernels were measured from: area, perimeter, compactness (4*pi*A/P^2), length, width, asymmetry coefficient and length of kernel groove. Each of kernel variety has the 70 samples. Samples were randomly selected for the experiment. Kernel structure was detected using a soft X-ray technique. Also structure of seed data set can be seen in table 1.

Table 1.Structure of seeds data set.

Name	Kama	Rosa	Canadian
Area	70 Sample	70 Sample	70 Sample
Perimeter	70 Sample	70 Sample	70 Sample
Compactness	70 Sample	70 Sample	70 Sample
Length of Kernel	70 Sample	70 Sample	70 Sample
Width of Kernel	70 Sample	70 Sample	70 Sample
Asymmetry Coefficients	70 Sample	70 Sample	70 Sample
Length of Groove	70 Sample	70 Sample	70 Sample

Methods

Classification is the process of obtaining information about the unknown data from the known data. For this purpose, database named as training set is created by known data. Learning algorithm is applied to the database and model is obtained. The model is applied to test data and result is obtained. Flow diagram of classification can be seen in figure 1.

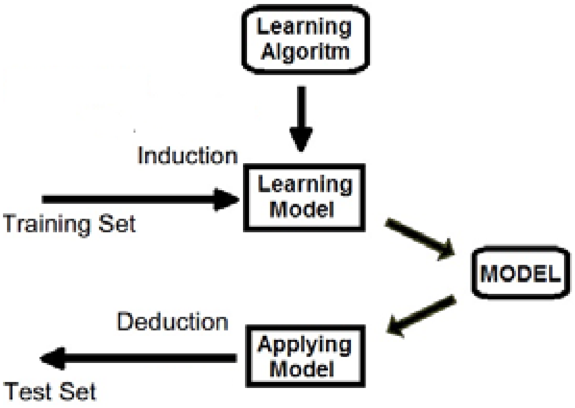


Fig. 1. Flow diagram of classification

The most widely used classification algorithms (Decision Trees, Neural Networks, Naive Bayes and K-NN) have been compared in this study (Kotsiantis, 2007; Xindong, et. al., 2008). Also in the similar studies related to seed classification only some of these methods (ANN and Naive-Bayes) were used and others were neglected.Logic of these methods are shortly explained in the below.

Support vector machines: includes test element to appropriate class based on hyper-plane created by known class labels. The primary purpose of this method is to determine the hyper-plane that makes maximizing the margin between classes.Thenmemberships of test elements are determined by the relative location to hyper-plane. Training stage consists of the creation of the hyper-plane (Cortes, et. al., 1995).

K-Nearest Neighbor: Test elements are included to classes according to number of k-nearest neighbor. The k-parameter is used to determine the number of nearest neighbors of test element. Test elements are classified by a majority vote of its "k" number of neighbors. For example, if k=1, then the object is simply assigned to the class of that single nearest neighbor. In this method there is no training. However, the choice of k parameter is the most important problem (Altman, 1992).

Naive Bayes: This method is based on the probability of belonging to the classes of elements. Test element is become member to class having highest probability of belonging (Hand, et al., 2001).

Decision Tree: In this method initially all the learning set samples are split recursively depending on the quantity of the selected root samplesAll of the samples belonging to the same class label is terminated as the leaf node. If not, the most appropriate samples will split the class attributes are selected (Quinlan, 1987).

For all methods, %90 of seeds data set was used for training and %10 of was used for testing.So, dataset is partitioned to 10 (%10 is for test and %90 is for train. Also ten-fold cross validation technique was used in order to obtain more accurate results.Thus all of the data were used for training and testing in each of ten cycles.

Results

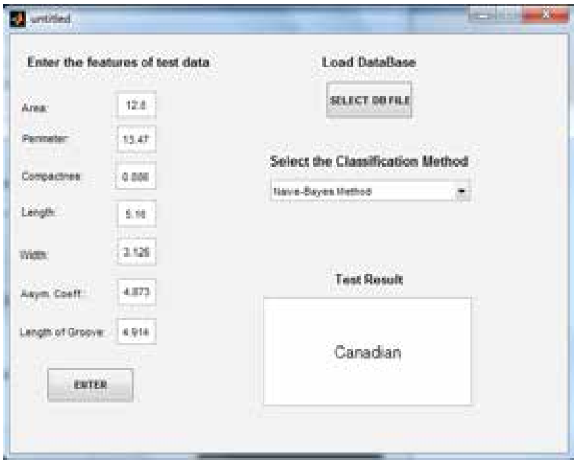
There are several classification methods in the literature. But, the most widely used classification algorithms (Decision Trees, SVM, Naïve Bayes and K-NN) have been compared in this study.Each of them has advantages and disadvantages. These advantages and disadvantages become important according to user purpose.In the field of seed classification the number of seed data to be classified will be more.Since the aim of this study is to determine the high applicable classification method in real life, the speed and accuracy rate of methods will be moreimportant.Therefore, primarily, classification methods need to extract based on the accuracy rateandspeed. In this study, well known classification methods were applied to the seed data set for evaluating the speed and accuracy rate together. Results can be seen in table 2.

Table 2.Experimental Results.

Methods	SVM	K-NN (11)	Naive-Bayes	Decision Tree
Accuracy Rate (%)	90,78	90,76	90,75	82,76
CPU Time (Second)	0,71	0,04	0,69	2,7

The value in parentheses is k-parameter that provides the highest performance to k-NN method.

According to results, accuracy rates of SVM, K-NN and Naive Bayes methods are approximately same. Consequently it can be said that other method is not appropriate in terms of accuracy rate. From the point of time of K-NN method can be considered as the best. But accuracy rate of this method has been obtained by optimum k-parameter. This method must be repeatedly tried until find the optimum k-parameter. This trial process needs more processing time. Therefore it is not possible to determine the availability of K-NN method in terms of time. Thus SVM and Naive-Bayes Methods can be said to be most suitable for seed classification. In this study the MATLAB software suitable for this purpose was performed. Image of this software can be seen in figure 2.


untitled
Enter the features of test data
Load DataBase
SELECT DB FILE
Select the Classification Method
Test Result
ENTER

Area: 12.8

Perimeter: 13.47

Compactnes: 0.886

Length: 5.16

Width: 3.126

Asym. Coeff: 4.873

Length of Groove: 4.914

Naive-Bayes Method

Canadian

Fig. 2. Proposed Software

As can be seen from figure 2 the seed will be tested can be determined by entering geometric features of seed.

Conclusion

In this study most suitable classification methods have been proposed to classification of seeds for use in real life. According to the experimental results both of the suggested methods are available for real life applications. For this purpose, the necessary software must be in the following way:

Database comprising the seed qualities and varieties must be created by expert agronomist. Also database must be obtained by image processing techniques since obtaining the geometric attributes is easier than obtaining other attributes.

New data to be tested should be obtained with the same image processing techniques.

Software must be used the proposed classification methods in this study.

In this study the software suitable for this purpose was performed and this software is proposed the aim of this study. When this software is loaded into the camera with image processing software it can be applicable in the real-life.

REFERENCES

Altman, N. S., (1992). An introduction to kernel and nearest-neighbor nonparametric regression.The American Statistician, 46(3), 175–185. | Charytanowicz, M. Niewczas, J. Kulczycki, P. Kowalski, P.A. and Lukasik, S. Zak, S. (2010). A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images.Information Technologies in Biomedicine, EwaPietka, JacekKawa, Springer-Verlag, Berlin-Heidelberg, 15-24. | Cortes, C. and Vapnik, V. (1995).Support-vector networks.Machine Learning 20 (3): 273-297. | Granitto, P.M. Navone, H. D. Verdes, P.F. and Ceccatto, H.A. (2002). Weed seeds identification by machine vision. Computers and Electronics in Agriculture, 33, 91–103. | Granitto, P.M. Verdes, P.F. andCeccatto, H. Alejandro, (2005). Large-scaleinvestigation of weedseedidentificationbymachinevision. ComputersandElectronics in Agriculture, 47, 15–24. | Hand, D. J. and Yu, K., (2001). Idiot's Bayes not so stupid after all. International Statistical Review, 69(3), 385–399. | Frank, A. and Asuncion, A., (2014). UCI Machine Learning Repository.http://www.archive.ics.uci.edu/ml. | Kotsiantis, S. B., (2007). Supervised Machine Learning: A Review of Classification Techniques.Informatica, 31, 249-268. | Luo, X. Jayas, D. S. Symons, and S. J. (1999).Identification of Damaged Kernels in Wheat using a Colour Machine Vision System.Journal of Cereal Science, 30, 49–59. | Quinlan, J. R. (1987). Simplifying decision trees.International Journal of Man-Machine Studies, 27(3), 221-234. | Wu, X. Kumar, V., Quinlan, J.R., Yang, Q., Hiroshi, M.,•McLachlan, G.J. Angus, N. Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J. nad Steinberg, D. (2008). Top 10 algorithms in data mining.Knowledge Information Systems, 14, 1-37. | Zhao-yan, L. Fang, C. Yi-bin, Y. Xiu-qin, RAO. (2005). Identification of riceseedvarietiesusingneural network. Journal of ZhejiangUniversityScience, 6(11), 1095-1100. |