Computers and Electronics in Agriculture 96 (2013) 173–179

Contents lists available at SciVerse ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires

José R. Romero [a], Pablo F. Roncallo [a,b], Pavan C. Akkiraju [a,b], Ignacio Ponzoni [c,d], Viviana C. Echenique [a,b], Jessica A. Carballido [a,*]

[a] CERZOS-CONICET, CCT-Bahía Blanca, Argentina
[b] Departamento de Agronomía, UNS, Argentina
[c] Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), DCIC, UNS, Argentina
[d] Planta Piloto de Ingeniería Química (PLAPIQUI-CONICET), CCT-Bahía Blanca, Bahía Blanca, Bs. As., Argentina

ARTICLE INFO

Article history:
Received 5 December 2012
Received in revised form 20 May 2013
Accepted 21 May 2013

Keywords:
Machine learning
Expert system
Classification algorithm
Yield
Triticum turgidum
Pasta wheat

ABSTRACT

Wheat is one of the most important cereals worldwide for human nutrition. Tetraploid wheat (*Triticum turgidum* L. ssp. *durum*, $2n = 28$, genomes AABB) is mainly used to produce pasta. The main objective of durum wheat breeding programs is to develop varieties with good quality and high yields. Yield is a very complex trait, and depends on different yield components that are genetically controlled and affected by environmental constraints. In this context, machine learning constitutes an excellent alternative for the analysis of a high number of traits in order to extract the most relevant ones as confident predictors of the performance of this crop, allowing a better agricultural planning. Thus, we propose the use of machine learning algorithms for the classification of yield components and for the search of new rules to infer high yields at harvest of durum wheat. The main objective of this work was to obtain rules for predicting durum wheat yield through different machine learning algorithms, and compare them to detect the one that best fits the model. In order to achieve this goal, One-R, J48, Ibk and A priori algorithms were run with data collected by our research group of a RIL (recombinant inbreed lines) population growing in six different environments from the Province of Buenos Aires in Argentina. The results indicate that the A priori method obtains the best performance for all locations, and the classificators generated using the different algorithms share a common set of selected traits. Moreover, comparing these results with the previous ones obtained using different techniques, mainly QTL mapping, the traits indicated to be the most significant ones were the same. The analysis of the resulting rules shows the soundness in the agronomic relevance of the extracted knowledge.

1. Introduction

An expert system is a computer program that simulates the judgment and behavior of a human that has expert knowledge and experience in a particular field. In this context, some classification algorithms from artificial intelligence try to extract expert knowledge, given a sufficient quantity and quality of data together with an adequate statistical and domain validation. In particular, the main objective within machine learning is to develop techniques that allow computers to automatically "learn" by means of generalizing behaviors from unstructured information provided in the form of examples that will be useful to create these generalizations/associations.

The software WEKA (Hall Mark, 2003) contains multiple machine learning algorithms for the implementation of supervised and unsupervised techniques. It is platform free (can be run on any operating system) and contains an extensive collection of methods for data preprocessing and modeling. Using these types of software systems with the appropriate data allows the finding of patterns that are able to predict association rules among different variables.

In agriculture, yield is one of the most important goals, and early assessment of yield reductions can prevent a disastrous situation and help in strategic planning to meet demands. Wheat is a major renewable resource for food, feed and industrial raw materials, and is among the major crops grown on the largest area worldwide. It is also one of the earliest crops to have grown on large scale, due to its high yields and to the possibility of long-term storage. The actual rate of wheat production increase (0.54% per year between 1997 and 2007) is less than half of that required in the next future (1.32% annual increase). As the area cropped with wheat may only marginally increase, further production must be mainly achieved by increasing yield (Reynolds et al., 2009). Durum

* Corresponding author. Tel.: +54 2914268797.
E-mail address: jac@cs.uns.edu.ar (J.A. Carballido).

0168-1699/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compag.2013.05.006

wheat (*Triticum turgidum*) or pasta wheat compared with common bread wheat (*Triticum aestivum*) is known for its hardness, protein content, intense yellow colour, nutty flavor and excellent cooking qualities. For these reasons it constitutes an excellent option for pasta production. This crop has also great importance in grain-producing areas of the Mediterranean and North America. The annual area planted worldwide is estimated at 13.5 million hectares, has shown a downward trend since 1970, since 18 million (Belaid, 2000). According to the "Integrated System of Agricultural Information" (*SIIA*),[1] Argentina increased the cultivated area with durum wheat from 52,420 ha in the period 2010/11 to 64,200 ha in 2011/12. The main area dedicated to this crop is the Province of Buenos Aires.

Globally, the main objective of durum wheat breeding programs is to develop varieties with high yields. Yield potential has been defined as "the yield of a cultivar when grown in environments to which it is adapted; with nutrients and water non-limiting; and with pests, diseases, weeds, lodging, and other stresses effectively controlled" (Evans and Fischer, 1999). However, in natural conditions different situations affect crop yields and it is important to have an idea of the performance of a variety in multiple environments to learn basic rules about it. Generally, grain yield is considered as a combination of the number and weight of the grains (Kuchel et al., 2007). Wheat grain yield can also be separated into its components, including spike components (thousand grain weight *Tgw*, grain number per year *Gne*, grain weight per year *Gwe*, spikelets number per year *Sne* and spike fertility *SF*) and can be correlated with agronomic (plant height Ph and harvest index HI) and morphological traits (peduncle length Pd). According to Cuthbert Janice et al. (2008), the analysis of these components along the growing period would allow to get knowledge about the genetic control and the relationship between yield and its components.

In this context, in order to select the most relevant traits for the prediction of durum wheat yield, the need of using machine learning methodologies arouse (Witten et al., 2005). This decision was made since it has been proven that these methods are robust and efficient in feature selection problems (Blum and Langley, 1997). Increasing demands for dimensionality reduction has broadly expanded research on feature selection into many fields, including pattern recognition, machine learning, data mining, and knowledge discovery. Feature selection, as a preprocessing step to machine learning, has proven to be very effective in reducing dimensionality, removing irrelevant data and improving comprehensibility (Langley et al., 1994; Liu and Motoda, 1998), and it has been used in different fields, like mechanics (Casimira et al., 2006), medicine (Palmerini et al., 2011), chemistry (Meydan and Sezerman, 2010) or chemioinformatics (Soto et al., 2009a).

In particular, several authors proposed the use of machine learning in agronomy (Boissard et al., 2008; Niedziela et al., 2012 and Atas et al., 2012). Boissard et al. (2008) remarked that the greatest challenge in horticulture is based on an early detection of diseases and on the reduction in the use of pesticides. These authors were able to create a working prototype that performs faster than a normal method for disease inspection. Niedziela et al. (2012) indicated that crop production practices and industrialization processes may result in acidification of arable soils. Associations between markers and traits were tested using a multiple linear model, as well as a statistical machine learning approach. With the use of this approach they were able to suggest the putative location of markers responsible for aluminum tolerance. Atas et al. (2012) used a new approach for aflotaxin detection in chili pepper. These authors have shown that the use of hyper spectral imaging and data mining techniques exhibits higher performance for aflotaxin detection in chili pepper than traditional methods. Therefore, the use of machine learning offers great promise to provide predictive models for processes involving agricultural systems.

In this paper, the impact of using machine learning techniques in predicting yield of durum wheat from its components is explored. The main objective of this work is to obtain rules for predicting durum wheat yield in early stages of crop development through different machine learning algorithms, and compare them according to different performance measures in order to assess the true potentiality of machine learning in this particular application. Moreover, we would like to establish a set of correlation-rules specific for the southeast of Buenos Aires Province, which can contribute to improve the yield of Durum wheat in the region. The article is organized as follows: at the beginning of Section 2 the platform and the algorithms used to carry out the studies are described, and then details about collection and preparation of the experimental data are presented, together with the evaluated traits. In Section 3, the results are presented and discussed. Finally, conclusions and some proposals are put forward in Section 4.

2. Materials and methods

The tool WEKA (http://www.cs.waikato.ac.nz/~ml/Weka/) was used for the generation of the predictive models. Weka is an open-source Java application produced by the University of Waikato, New Zealand. This software is available for free on the official site of the institution and contains multiple algorithms for the implementation of supervised and unsupervised techniques. The software also includes an extensive collection of techniques for data preprocessing and modeling, providing a friendly interface for training and validation of models (Hall Mark, 2003).

2.1. Algorithms

In this section a brief review about the machine learning algorithms that were used in this work is presented. The algorithms were selected based on the previous experience of the authors, and considering their aptitude to solve problems with the following features: classification and association of variables in discrete categories, in terms of the volume of data, so that the different categories help in the prediction of yield of durum wheat. Table SM1 summarize the main features of the algorithms used to asses yield and yield components in Durum wheat.

The **One-R** algorithm, short for "One Rule", is a simple, yet accurate, classification algorithm that generates a one-level decision tree (Shi, 2007). It creates one rule for each attribute in the training data, and then selects the rule with the smallest error rate as its "one rule". To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is the class that appears most often for that attribute value. A rule is a set of attribute values bound to their majority class. In WEKA, this algorithm selects the rule with the highest number of correct instances and not the one with the least error rate (Mitchell, 1997).

The **J48** algorithm (Weka implementation of C4.5) is also part of the algorithms based on decision trees, like the previous one (Ronny Kohavi, 2002). The key feature of this algorithm is that it incorporates a classification tree pruning once it has been induced, i.e., when the decision tree is built, the branches with less predictive power are pruned. This algorithm is an enhancement of ID3 (Quinlan, 1986), also based in trees, where the criteria chosen to select the most informative variable are based on the concept of

[1] SIIA: Sistema Integrado de Información Agropecuaria (http://www.siia.gov.ar).

amount of mutual information between this variable and the variable class (Witten et al., 2005, pp. 404–410).

The **IBK** algorithm is based on instances; therefore, it consists only in storing the data presented. When a new instance is found, a set of similar related instances is returned from memory, and is used to classify the consulted instance (Aha et al., 1991). It is, therefore, a lazy learning algorithm. This learning method is based on the classification modules that keep in memory a selection of examples without creating any kind of abstraction in the form of rules or decision trees (hence the name, lazy) (Kamber, 2006). Every time a new instance is found, calculate its relation to the examples previously saved for the purpose of assigning a value to the objective function for the new instance. The idea is that a new case should be classified as the most frequent class when it belongs to its K nearest neighbors, being assigned to the class most common amongst its k nearest neighbors measured by a distance function. Hence, it is also known as KNN method (K Nearest Neighbors) (Lu, 2011).

Apriori is a classic algorithm used to find association rules in a dataset, widely used for discovering relationships between variables in large databases (Piatetsky-Shapiro, 1991). This algorithm is based on prior knowledge or "a priori" given a set of itemsets, where the itemsets represent putative association rules between problem variables. Basically, the algorithm attempts to find subsets which are common to at least a minimum number of the item sets, it's a bottom search, moving upward level; it prunes many of the sets which are unlikely to be frequent sets, reducing the search space and increasing the efficiency (Agrawal and Srikant, 1994; Härdle Wolfgang, 2002).

For more in deep details about the mentioned methods see Han and Kamber (2000).

2.2. Biological datasets

The data used in this work were taken from the PhD thesis of Pavan Chand Akkiraju related with the mapping of genomic regions associated with yield and yield components in durum wheat within different environments (Akkiraju, 2010).

2.2.1. Plant Material

A mapping population consisting of 93 recombinant inbred lines (RILs) was obtained by crossing the line UC1113 with the variety Kofa (Zhang et al., 2008). UC1113, a breeding line from the UC Davis Wheat Breeding Program, has excellent agronomic performance, but intermediate pasta quality parameters. Kofa is a durum variety developed by West-Bred. It has optimal semolina and pasta color, high protein content and strong gluten.

2.2.2. Experimental design and field trials

The experiments were conducted and evaluated in three locations of the Province of Buenos Aires, Argentina, during two growing seasons (2006/2007 and 2007/2008). The locations were Cabildo (39°36′S61°64′W), Barrow (38°20′S60°13W), and Balcarce (37°45′S58°18′W). The rainfall recorded in 2006 between August and December in these locations was 175.4 mm, 344.6 mm and 271.2 mm, respectively. The corresponding rainfall values for 2007 were 233.2 mm, 286.9 mm, 366.5 mm for Cabildo, Barrow and Balcarce, respectively.

The experimental design was a randomized complete block (RCBD) with three replications, using 3 m^2 plots. At each location the trial received special care and fertilization management, appropriate to the area where the experiments were conducted. The RILs population and their parents (UC1113 and Kofa) were evaluated in the mentioned conditions and locations.

2.2.2.1. Evaluated traits. Two kinds of data were used, plot data and individual plant data, as follows:

(a) Plot data

Grain yield (Yld) from each entire plot was obtained by weighing the clean grains harvested using a harvest machine (Kg/ha). Thousand grain weight (Tgw) was recorded by weighing two samples of 100 grains from each plot. Each value was used to calculate thousand grains weight and then averaged, expressed in grams (g).

(b) Individual plant data

Ten plants from each RIL were randomly collected from the central row of each plot after harvest maturity. The following yield related traits were measured on each plant: plant height (cm), peduncle length (cm), harvest index, spikelets number per year, grain number per year, grain weight per year, and spike fertility. Average values were calculated per plot by using the data of the ten plants.

- *Plant height (Ph)*: Calculated as the distance from the edge of separation of the stem from the root to the tip of the spike, and it was expressed in cm.
- *Peduncle length (Pd)*: Measured as the distance from the last internode to the base of the spike, expressed in cm, and obtained as the average of the measurements in all tillers from each plant.
- *Harvest index (Hi)*: Quantified as the ratio between the total weight of grains per plant (Wgp), and the weight of the plant (WP). (Hi = Wgp/Wp).
- *Spikelet number/ear (Sne)*: Obtained as the average number of spikelets/ear, counting the number of spikelets in all the ears/plant.
- *Grain number/ear (Gne)*: Calculated as the product of the weight of grains/spike (Gwe) and the weight of one grain which was obtained from the thousand grain weight (Tgw) (Gne = Gwe × (1000/Tgw).
- *Grain weight/ear (Gwe)*: Determined by weighing the grains from each ear of the plant. The value per plant was calculated as an average of all ears by plant.
- *Spike fertility (SF)*: Calculated as the ratio of the number of fertile spikelets/ear (Nfse) and the number of total spikelets/ear (Ntse), and expressed as a percentage (SF = (Nfse)/(Ntse) ∗ 100).

2.3. Data preparation for analysis using the Weka software

The data obtained from agronomic experiments were submitted to a comma delimited CSV file with 8 records (traits analyzed) dismissing RIL's column. As all input data were real numbers, it was necessary to discretize them using an unsupervised filter in order to make feasible the use of classification algorithms that do not handle these numeric attributes. Ten boxes (bins) of possible intervals were used for each trait. The boxes represent ranges: they arise from the need to discretize in sub-ranges the real numbers that are associated to the attributes.

In order to classify the yield components and relate them to the real yield (calculated as Kg/plot) of each RIL/location, the actual values were divided into three groups corresponding to low, medium and high yield. This discretization into classes depended of the specific agronomic region. For Cabildo the values were: less than 899 kg s for low, 900–1799 kg s for medium and greater than 1800 kg s for high; Barrow: less than 2132 kg s for low, 2133–4264 kg s for medium and greater than 4264 kg s for high; Balcarce: less than 1476 kg s for low, 1477–2952 kg s for medium and greater than 2953 kg s for high); and the combined location:

less than 2132 kg for low, 2133–4264 kg for medium and greater than 4264 for high. In this context, it is important to have in mind that the objective of this work is to show a global analysis, being aware of the limitations of realizing a join of the data. Each data set for each location was divided (80/20): the first (80% of the data) was used as the training set and the second (20%) was used for the evaluation; this division was performed under an instance supervised filter called "Stratified-RemoveFolds" by Weka.

3. Results and discussion

As it was described in Material and Methods, the three locations had different environmental conditions that affect in different ways the yield of this crop. Balcarce provides the best environment for obtaining good yield values, Cabildo is the worst, having Barrow intermediate values. Results obtained after applying the algorithms to the durum wheat field datasets are shown here, described in the following paragraphs and summarized in Table 1. Table SM2 shows the traits that intervene in the rules that correspond to each location.

Regarding to the accuracy level, the **Apriori** algorithm was identified as the most relevant for each location and for the combination of the three. Accuracy, often called confidence, is the number of instances that it predicts correctly, expressed as a proportion of all instances to which it applies. In our context, the accuracy represents the proportion of samples with which the algorithms predict the correct yield class (high, medium and low yield).

3.1. One-R

Using this algorithm it was possible to associate low, medium and high yield to plant height in Balcarce, peduncle length with high, medium and low yields in Barrow and spike fertility with high and medium yields in Cabildo. These predictions were obtained with accuracies of 57.00%, 65.14% and 93.20%, respectively. When the values from all the three combined locations were considered it was possible to predict high and medium yields from the variable thousand grain weight with an accuracy of 74.90%.

In summary, with the One-R algorithm it was possible to detect three different yield predictors for all the three locations, but with a low level of accuracy except for the predictor based on spike fertility in Cabildo. Plant height, peduncle length and spike fertility are interesting traits since they can be measured in a pre-harvest stage, when the plant reaches its maximum development. These data allow the anticipation of the final yield of the crop, with applications in plant breeding programs or harvest prediction at farmer scale. Thousand grain weight appears as a good indicator to predict high yields with high accuracy in the global analysis (using all datasets together), and as it is already known, is a useful trait traditionally used in plant breeding.

3.2. J48

In Balcarce, the use of this algorithm to predict high yield based on thousand grain weight, peduncle length, harvest index, grain number per year, grain weight per year, plant height, spikelet number per year and spike fertility, gave a final accuracy level of 71.03%. In Barrow this algorithm allowed the prediction of high yield based on six characteristics: thousand grain weights, plant height, spikelet number per year, grain number per year, spike fertility, and peduncle length with a final accuracy level of 66.97%.

Using this algorithm it was not possible to predict yield for Cabildo since this location had adverse characteristics for growing this cereal in the evaluated period (lower temperatures and lower rainfall than in Balcarce or Barrow, especially in 2006). The low variation among the data for this location makes the tree collapse in a single node. For the algorithm J48, if all the records consist in the same value for the target attribute, it returns a single leaf node with that value. With this algorithm for the three combined locations it was possible to determine that the most important traits to predict high yields were thousand grain weight, spikelet number per year, spike fertility, grain number per year, plant height, harvest index, peduncle length and grain weight per year with a final accuracy level of 76.78%.

Table 1
Accuracy level in predicting high yields of durum wheat, using different algorithms for three locations of the province of Buenos Aires.

	One-R (%)	J48 (%)	IBK (%)	Apriori (%)
Balcarce	57.00	71.03	71.03	76.00
Barrow	65.14	66.97	61.47	79.00
Cabildo	93.20	92.23*	93.20	96.00
Three combined locations	74.90	76.78	77.15	90.00

* The low variation among the data for Cabildo makes the tree collapse in a single node.

3.3. IBK

The absence of a universal optimal value of k, valid for any problem, there was to experiment with values ($K = 2,3,4,5$) and determine the best value that produces better accuracy of the nearest neighbor algorithm. The value that best fitted our data was $k = 2$. Using this algorithm with data from Balcarce a level of accuracy of 71.03% was found for the traits showed in Table 1. For Barrow, the accuracy level was 61.47% whereas the accuracy level for the traits measured in Cabildo was higher, explaining the model with an accuracy of 93.20%. Similar results were obtained for the three combined locations (accuracy level 77.15%). Table 2 shows the continue-valued ranges corresponding to the different traits used for predicting high yields using the IBK algorithm, when the real yields values (measured as it was indicated before) are high.

3.4. Apriori

Using this algorithm it was possible to associate medium yield with plant height, with values between 81.37 and 85.68 cm, in Balcarce. In Barrow the association was a medium yield with peduncle length, with values between 27.01 and 29.23 cm. These predictions had accuracies of 76.00% and 79.00%, respectively. However, in Cabildo the algorithm indicated that the best predictors for high yield were the following traits: thousand grain weight with values between 31.31 g and 33.38 g, plant height with values ranging 65.64 cm and 69.70 cm, and peduncle length with lengths between 24.5 cm and 28 cm. All these traits with a accuracy level of 96%. When the data of the three locations were combined, only one rule for predicting high yield was found. Thousand grain weight, with values between 29.66 cm and 33.85 cm, showed an accuracy of 90%. Two important conclusions can be drawn from this information; one is the high level of accuracy, not only for the individual locations but also for the combined locations. Another one is the coincidence between this algorithm and One-R to detect thousand grain weight as a good indicator of high yield for the three combined locations. Table 3 shows the location and the rules associated with yield in each location.

Therefore, our study indicates that the algorithms that performed with the highest accuracy are j48 an Apriori. There is a good reason for preferring simpler models; they are easier for people to understand, remember and use (as well as cheaper for computers to store and manipulate). Given two models with the same accuracy, the simpler on should be preferred because simplicity is desirable in itself. Considering the Occam's principle of parsimony or Occam's razor that states that one should always

Table 2
Ranges associated to the traits used for predicting high yields using the IBK algorithm.

		Balcarce	Barrow	Cabildo	Three combined locations
Tgw	Thousand grain weight	{38.98,60.46}	{41.34,48.02}	{25.07,+∞}	{−∞,42.22} U {46.40,54.78}
Ph	Plant height	{−∞,55.52} U {59.83,81.37} U {85.68,v89.98}	{77.4,81.05}	{−∞,53.46} U {57.52,85.94}	{−∞,89.80}
Pd	Peduncle length	{−∞,38.32}	{31.45,33.68}	{−∞,34.5}	{−∞,37.64}
Hi	Harvest index	{−∞,0.29} u {0.34,0.53}	{0.38,0.44}	{0.21,0.46}	{0.28,0.46}
Sne	Spikelet number/ear	{−∞,11.74} U {12.48,17.66}	{15.14,15.64} U {16.15,16.66}	{11.23,17.80}	–
Gne	Grain number/ear	{10.49,38.66}	{23.12,26.40} U {32.94,36.12}	–	{22.84,39.84}
Gwe	Grain weight/ear	{0.49,1.80}	{1.1,1.21} U {1.56,1.67}	{−∞,1.56}	{−∞,1.66}
Sf	Spike Fertility	{0.33,0.70} U {0.78,+∞}	{0.75,0.79} U {0.83,0.87}	{0.51,0.57} U {0.69,+∞}	{0.63,+∞}

Meaning any value.

Table 3
Best rules associated with durum wheat yields in different locations from the province of Buenos Aires using Apriori algorithm.

Location	Rule
Balcarce	Ph = ′(81.369–85.676]′==>Category = Medium conf:(0.76)
Barrow	Pd = ′(27.01–29.2′2]′ ==> Category = Medium conf:(0.79)
Cabildo	Tgw = ′(31.308–33.3′5]′==> Category = High conf:(0.97)
	Ph = ′(65.64–69′7]′==> Category = High conf:(0.96)
	Pd = ′(24.5–27]′==> Category = High conf:(0.96)
All together	Tgw = ′(29.662–33.8′8]′==>Category = High conf:(0.9)

opt for an explanation in terms of the fewest possible causes, factors or variables, the simplicity is a goal in itself, we should choose Apriori algorithm. This algorithm proved to be the one that finds the simplest rule to explain high yield for each location.

In summary, the results of this work indicate that it is possible to use different algorithms working under the software Weka to predict yield in different locations from the Province of Buenos Aires in Argentina. The data of two working years were used to create a system able to predict yield in different stages of plant development. Providing more data to the system (from 3 to 4 years) could give more precision in the results. However, comparing these results with the previous ones from our group (Akkiraju, 2010) and with results from other researchers (Maccaferri et al., 2008, 2010) using different techniques, mainly QTL mapping, the traits indicated as more significant were the same. Table 4 shows the times that each trait was implicated in high yield prediction in the durum wheat RIL population used for this study through specific algorithms. To establish which traits are involved with high yield for a specific algorithm, the number of times that the algorithms refers to these traits were determined, and represented this with a star. The number of stars is shown in Table 4. For example the algorithm Apriori has the trait Tgw involved twice, and one for Ph and Pd.

As it was mentioned before, they are the same traits as the ones reported in several articles by different authors. In particular, Akkiraju reported 74 significant QTL for yield and its components, especially for thousand grain weight (8), peduncle length (8), grain number/total spikelets (8), plant height (7), grain number/ear (7) and spike fertility (6). Several of these QTLs were stable and distributed on different wheat chromosomes. Moreover, peduncle length has been already mentioned by Maccaferri et al. (2008, 2010) as mapping on a region of chromosome 3B and strongly involved in conferring high yields to durum wheat.

Moreover, most yield variations are associated with those in grain number, both under different environments (Fischer, 1985; Savin and Slafer, 1991; Magrin et al., 1993) and as a result of genetic gains in yield potential (Slafer et al., 1990; Fischer, 2007). The number of fertile flowers at anthesis essentially determines grain number (Gonzalez Fernanda et al., 2011). The increased number of grains per spike was the main determinant of improved grain number through breeding in wheat (Siddique et al., 1989; Slafer et al., 1990; Slafer and Andrade, 1993; Calderini et al., 1999). Therefore, the rules extracted for the machine learning methods are consistent with previous biology knowledge, showing the soundness of these approaches. On the other hand, it is important to note that any prediction model obtained by machine learning methods can usually suffer of generalization problem. This issue refers to the ability of an algorithm to perform accurately on new, unseen examples after having been trained on a learning data set. Therefore, in our context, there are no guaranties about 379 the effectiveness of our yield level predictors if they were applied in environments with different features of those of the Province of Buenos Aires, which constitutes a limitation of these machine learning approaches.

Nevertheless, the impact of the generalization problem can be reduced by using applicability domain techniques. These

Table 4
Yield components measured in durum wheat Ril population and different algorithms used to obtain rules for predicting high yield. Stars (*) indicate how many times each trait was selected in high yield prediction through specific algorithms.

	Traits	Algorithms				#
		One-R	Apriori	j48	Ibk	
Spike components	Tgw	*	**	***	****	10
	Gne			***	***	6
	Gwe			**	****	6
	Sne			***	***	6
	Sf	*		***	****	8
Agronomic correlated	Ph	*	*	***	****	9
	Hi			**	****	6
Morphological traits	Pd	*	*	***	****	9

methods - commonly used in the field of quantitative structure – activity relationship models (QSAR models) [Jaworska et al., 2005; Grammatica, 2007] - use knowledge and information about the training set used to generate a predictor in order to identify its applicability range. Then, as we have experience in the development of this kind of techniques in the area of QSAR [Soto et al. 2009b, Soto et al. 2011], we expect to adapt these strategies to this agronomical field as future work.

4. Conclusions

In this paper, the use of machine learning strategies for the prediction of wheat yield from several phenotypic plant traits, corresponding to different locations of the Province of Buenos Aires in Argentina, was explored. Several classification methods were selected and tested considering their aptitude for extracting association rules from discrete-categorized target variables. As a part of this problem, the selection of the most relevant phenotypic plant traits for the prediction of yield arose. In this way, a double contribution was pursued in this work. First, we evaluated the power of machine learning methods as prediction tools in this context. Second, as a consequence of the inference of prediction models, a relevant analysis about the different traits with respect to the wheat yield levels was also achieved.

The results revealed that the A priori method outperformed the other tested techniques for all locations, but the set of selected traits were shared for the predictors (classifiers) obtained by the different algorithms. Regarding this point, it is interesting to remark that the traits selected as relevant by the machine learning methods were consistent with previous results obtained using other technologies, like QTL mapping. Moreover, the analysis of the resulting rules showed the soundness in the agronomic relevance of the extracted knowledge.

Acknowledgements

This work has been funded by Grants PIP112-2009-0100322 founded by the CONICET (National Research Council of Argentina), PICT1011 from ANPCyT and by Grants PGI 24/ZN15, PGI 24/ZN21, PGI-TIR 24/A185 founded by the Universidad Nacional del Sur (Bahía Blanca, Argentina).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compag.2013.05.006.

References

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 478–499.

Aha, W. David, Kibler, Dennis, Albert, Marc K., 1991. Instance-based learning algorithms. Machine Learning 6, 37–66.

Akkiraju, Pavan Chand, 2010. Mapeo de regiones genómicas determinantes del rendimiento y sus componentes en *Triticum turgidum* L. var. durum en diferentes ambientes. Universidad Nacional del Sur. Departamento de Agronomía, 2010. xxvii, 211 h.

Atas, M., Yardımcı, Y., Temizel, A., 2012. A new approach to aflatoxin detection in chili pepper by machine vision. Computers and Electronics in Agriculture 87 (C), 129–141.

Belaid, A., 2000. Durum wheat improvement in the mediterranean region: new challenges. In: Belaid, A. (Ed.), Durum Wheat Improvement in the Mediterranean Region: New Challenges. Options Méditerraneens, CIHEAM, Zaragoza, pp. 33–50.

Blum, Avrim L., Langley, Pat, 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271.

Boissard, Paul, Martin, Vincent, Moisan, Sabine, 2008. A cognitive vision approach to early pest detection in greenhouse crops. Computers and Electronics in Agriculture 62 (2), 81–93.

Calderini, D.F., Reynolds, M.P., Slafer, G.A., 1999. Genetic gains in wheat yield and main physiological changes associated with them during the 20th century. In: Satorre, E.H., Slafer, G.A. (Eds.), Wheat: Ecology and Physiology of Yield Determination. Food Product Press, New York, pp. 351–377.

Casimira, R., Boutleuxa, E., Clercb, G., Yahouib, A., 2006. The use of feature selection and nearest neighbors rule for faults diganostic in induction motors. Engineering Applications of Artificial Intelligence 19 (2), 168–177.

Cuthbert Janice, L., Somers Daryl, J., Brûlé-Babel, Anita L., Brown Douglas, P., Crow Gary, H., 2008. Molecular mapping of quantitative trait loci for yield and yield components in spring wheat (*Triticum aestivum* L.). Theoretical and Applied Genetics 117 (4), 595–608.

Evans, L.T., Fischer, R.A., 1999. Yield potential: its definition, measurement and significance. Crop Science 39 (6), 1544–1551.

Fischer, R.A., 1985. Number of kernels in wheat crops and the influence of solar radiation and temperature. Journal of Agricultural Science 100, 447–461.

Fischer, R.A., 2007. Understanding the physiological basis of yield potential in wheat. Journal of Agricultural Science 145, 99–113.

Gonzalez Fernanda, G., Miralles Daniel, J., Slafer Gustavo, A., 2011. Wheat floret survival as related to pre-anthesis spike growth. Journal of Experimental Botany. http://dx.doi.org/10.1093/jxb/err182.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Combinatorial Science 26 (5), 694–701.

Hall Mark, E.F., 2003. The WEKA data mining software: an update. SIGKDD Explorations 11 (1).

Han, Jiawei, Kamber, Micheline, 2000. Data Mining: Concepts and Techniques. Simon Fraser University, Morgan Kaufmann Publishers.

Härdle Wolfgang, B.R., 2002. In: Proceedings in Computational Statistics. Compstat. Berlin, Germany, p. 648.

Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T., 2005. QSAR applicability domain estimation by projection of the training set descriptor space: a review. Alternatives to Laboratory Animals 33 (5), 445–459.

Kamber, M.H., 2006. Data Mining: Concepts and Techniques, second ed. Elsevier, Amsterdam.

Kuchel, H., Williams, K.J., Langridge, P., Eagles, H.A., Jefferies, S.P., 2007. Genetic dissection of grain yield in bread wheat. I. QTL analysis. Theoretical and Applied Genetics 115 (8), 1029–1041.

Langley et al., 1994. Selection of relevant features in machine learning. In: Proceedings of the AAAI Fall Symposium on Relevance.

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery Data Mining. Kluwer Academic Publisher, Boston.

Lu, H.H.-S., 2011. Handbook of Statistical Bioinformatics. Springer.

Maccaferri, M., Sanguineti, M.C., Corneti, S., Ortega, J.L., Ben Salem, M., Bort, J., DeAmbrogio, E., Garcia del Moral, L.F., Demontis, A., El-Ahmed, A., Maalouf, F., Machlab, H., Martos, M., Moragues, M., Motawaj, J., Nachit, M., Nserallah, N., Ouabbou, H., Royo, C., Slama, S., Tuberosa, R., 2008. Quantitative trait loci for grain yield and adaptation of durum wheat (Triticum durum Desf.). Across a Wide Range of Water Availability. Genetics 178, 489–511.

Maccaferri, M., Sanguineti, M.C., Demontis, A., El-Ahmed, A., Garcia del Moral, L., Maalouf, F., Nachit, M., Nserallah, N., Ouabbou, H., Rhouma, S., Royo, C., Villegas, D., Tuberosa, R., 2010. Association mapping in durum wheat grown across a broad range of water regimes. Journal of Experimental Botany. http://dx.doi.org/10.1093/jxb/erq287.

Magrin, G.O., Hall, A.J., Baldy, C., Grondona, M.O., 1993. Spatial and inter-annual variations in the photothermal quotient: implications for the potential kernel number of wheat crops in Argentina. Agriculture Forest Meteorology 67, 29–41.

Meydan, Cem., Sezerman, O. Uğur, 2010. Biomarker discovery for toxicity. Neurocomputing 73 (13–15), 2384–2393.

Mitchell, T.M., 1997. Sequential covering algorithms. In: Mitchell, T.M. (Ed.), Machine Learning. McGraw-Hill Science/Engineering/Math, pp. 275–282.

Niedziela, Agnieszka, Bednarek, Piotr T., Cichy, Henryk, Budzianowski, Grzegorz, Kilian, Andrzej, Anioł, Andrzej, 2012. Aluminium tolerance association mapping in triticale. BMC Genomics 13 (1) (Article number 67).

Palmerini, L., Rocchi, L., Mellone, S., Valzania, F., Chiari, L., 2011. Feature selection for accelerometer-based posture analysis in Parkinsons disease. IEEE Transactions of Information Technology in Biomedicine 15 (3), 481–490 (Article number 5719552).

Piatetsky-Shapiro, G., 1991. Discovery, Analysis, and Presentation of Strong Rules. In: Piatetsky-Shapiro, Gregory, Frawley, William J. (Eds.), Knowledge Discovery in Databases, pp. 229–260.

Quinlan, J.R., 1986. Induction of Decision Trees. Springer, Netherlands, vol. 1(1), pp. 81–106.

Reynolds, M., Foulkes, M.J., Slafer, G.A., Berry, P., Parry, M.A.J., Snape, J.W., Angus, W.J., 2009. Raising yield potential in wheat. Journal of Experimental Botany 60, 1899–1918.

Ronny Kohavi, J.R., 2002. Data mining tasks and methods: classification: decision-tree discovery. In: Ronny Kohavi, J.R. (Ed.), Handbook of Data Mining and Knowledge Discovery. Oxford University Press, New York, NY, USA, pp. 267–276.

Savin, R., Slafer, G.A., 1991. Shading effects on the yield of an Argentinean wheat cultivar. Journal of Agricultural Science 116, 1–7.

Shi, H., 2007. Best-First Decision Tree Learning. In: Shi, H. (Eds.), Master's thesis. University of Waikato, pp. 129–136.

Siddique, K.H.M., Kirby, E.J.M., Perry, M.W., 1989. Ear:stem ratio in old and modern wheat varieties; relationship with improvement in number of grains per ear and yield. Field Crops Research 21, 59–78.

Slafer, G.A., Andrade, F.H., 1993. Physiological attributes to the generation of grain yield in bread wheat cultivars released at different eras. Field Crops Research 31, 351–367.

Slafer, G.A., Andrade, F.H., Satorre, E.H., 1990. Genetic-improvement effects on pre-anthesis physiological attributes related to wheat grain yield. Field Crops Research 23, 255–263.

Soto, A.J., Cecchini, R.J., Vazquez, G.E., Ponzoni, I., 2009a. Multi-Objective Feature Selection in QSAR/ QSPR using a Machine Learning Approach. QSAR & Combinatorial Science 28, 1509–1523.

Soto, A.J., Ponzoni, I., Vazquez, G.E., 2009b. Segregating confident predictions of chemicals' properties for virtual screening of drugs. In: Omatu, Sigeru, Rocha, Miguel P., Bravo, Jose, Fernández, Florentino, Corchado, Emilio, Bustillo, Andres, Corchado, Juan M. (Eds.), Dist. Comp., Art. Int., Bioinformatics, Soft Comp.& Amb. Assisted Liv., IWANN 2009, vol. 2, Salamanca, Spain, June 10–12, Lecture Notes in Computer Science, vol. 5518. Springer-Verlag, Berlin Heidelberg, pp. 1005–1012.

Soto, A.J., Vazquez, G.E., Strickert, M., Ponzoni, I., 2011. Target-driven subspace mapping methods and their applicability domain estimation. Molecular Informatics 30, 779–789.

Witten, Ian H., Frank, Eibe, Hall, Mark A., 2005. Data Mining: Practical Machine Learning Tools and Techniques. In: E.F. Ian H. Witten. Elsevier.

Zhang, W., Chao, S., Manthey, F., Chicaiza, O., Brevis, J.C., Echenique, V., Dubcovsky, J., 2008. QTL analysis of pasta quality using a composite microsatellite and SNP map of durum wheat. Theoretical and Applied Genetics 117, 1361–1377.