Computers and Electronics in Agriculture 113 (2015) 136–147

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Data mining in the reduction of the number of places of experiments for plant cultivates

Wanderley Clarete Lanza Meirelles [a,1], Luis E. Zárate [b,*]

[a] Nucleus of Information Technology, Embrapa Maize and Sorghum, Road MG 424, km 65, Sete Lagoas, MG 35701-970, Brazil
[b] Department of Computer Science, Applied Computational Intelligence Laboratory – LICAP, Pontifical Catholic University of Minas Gerais, Av. Dom José Gaspar 500, Coração Eucarístico, Belo Horizonte, 30535-610 MG, Brazil

ARTICLE INFO

Article history:
Received 7 August 2014
Received in revised form 11 February 2015
Accepted 12 February 2015

Keywords:
Data mining
Maize cultivation
Clustering
Artificial neural networks

ABSTRACT

When a new plant cultivar is developed, it needs to be evaluated in relation to different environmental conditions before being commercially released. A set of experiments systematically was conducted at various locations with the purpose of evaluating the performance of the plants, in different soil, climate and elevation conditions. In these experiments, the plant culture development was monitored throughout the entire planting cycle. It is important to note that these experiments are very expensive and the aim of this work is to verify the possibility of reducing the number of places of the experiments without losing of representativeness in the evaluation of planted cultivars. To achieve this goal, the interaction among soil, climate and plant was modeled and the major components in each of these systems were considered. Neural computational models were also developed in order to estimate data regarding the relative air humidity. Using the partitioned clustering technique (k-medians), clusters (planting sites) were built in such a way that, according to the methodology proposed were similar in terms of soil, climate and plant behavior. The consistency of the formed clusters was evaluated by two criteria and the recommendation for reducing the cultivation sites was presented. By analyzing the results, it was possible to verify that the clustering process is not sufficient for deciding which planting locations should be omitted. It is necessary to count on complementary evaluation criteria associated with the production and cultivar ranking. In this work, the national competition experiments of maize cultivars in Brazil are considered as case study.

1. Introduction

The development of disease-resistant and better adapted plants to various regional conditions is an important condition to raise the productivity. When a new cultivar is developed, it needs to be planted and evaluated in relation to its behavior and its development in different environmental conditions, and thus verify the feasibility of its commercial launch. This evaluation is performed by a testing network (experiments) that is deployed in different regions obeying specific planting techniques. Then, the whole culture development is monitored and several variables are measured, generating a large data volume that is subsequently analyzed.

Considering large territorial extent with diversity of climate and soil conditions, a considerable amount of experiments in different locations is required to better express the interactions of cultivars with the environment. The higher the number of installed experimental sites, the better the description of this interaction. However, installing these sites is expensive, requiring trips of local teams to several planting locations, and constant monitoring throughout the crop cycle. Budget constraints also demand the reduction of the amount of installed experiments.

By means of a knowledge discovery process in databases (KDD) (Fayyad et al., 1996), the data generated from the experimental sites can be used to reduce the number of locations where the experiments are installed without losing the quality and reliability of the results. By means of clustering technique, this data would be used to identify regions with similar characteristics, thereby identifying "typical locations" that faithfully represent a region or a set of locations, then exempting the experimental facility in other places with similar characteristics. As a selection criteria, one may consider two or more similar sites if, after applying the clustering technique, the places belonging to the same cluster present the same harvest productivity.

* Corresponding author. Tel.: +55 31 3319 4117; fax: +55 31 3319 4001.
E-mail addresses: wanderley.lanza@embrapa.br (W.C.L. Meirelles), zarate@pucminas.br (L.E. Zárate).
[1] Tel.: +55 3130271912.

http://dx.doi.org/10.1016/j.compag.2015.02.006
0168-1699/

In order to achieve the goal of reducing the number of planting sites, first it is necessary to model conceptually the *plant–environment* system taking into consideration its main components and interactions. The aim is to identify the main variables involved in the system. The modeling must be based on as much explicit as in tacit knowledge. For this, interviews with experts in various fields such as soil, climate, plant physiology and breeding should be performed. Starting from the conceptual model, the needed data would be collected to be used by the KDD process and by the data clustering technique to achieve the aim of reducing the number of planting sites.

In this work, maize cultivars are considered as case study. The National Maize Competition Experiment in Brazil consists of a set of experiments conducted at several locations that follow the same planting procedures. In each experiment, the same types of cultivation (maize plant) are used and the whole crop development is monitored. Information is collected on grain yield, cob production, plant height, insect attack, diseases and various other data, forming a database that can be used to compare the performance of each cultivar in all locations where the experiments were performed. The implementation is done in a cooperative manner between public and private institutions under the coordination of the Embrapa Maize and Sorghum unit of the Brazilian Agricultural Research Corporation (Embrapa), with representatives of the Brazilian Seeds and Seedlings Association (Abrasem).

The data available for this work was consistent with the National Hybrid Elite Maize Experiment (advanced maize and with great potential for commercial release) from 2003/2004 crops involving 32 maize cultivars (plants) in 35 tests (experiments) conducted in 30 locations. It is important to note that some data (climate variables) were not available at the information system generated by the testing network, which led to the need for construction of computational models based on artificial neural networks (ANN) to estimate these values.

This work is organized as follows: Section 2 the characterization of the problem domain and a review on the usage of data mining techniques in agriculture are presented. In Section 3, the methodology applied and all stages taken to assembly the database are presented. In Section 4, the experimental results are analyzed. Finally, in Section 5, the conclusions of this work are presented.

2. Characterization of the problem domain and review of literature

The development of a plant can be described as the result of the interaction of three main factors: genotype, soil and climate. It was raised through literature and interviews with experts in soil, climate, plant physiology and breeding, which components influence the development of the maize plant. Table 1 shows the variables, plant, soil and climate considered for this work. Information concerning the plant genetic characteristics was not considered.

Taking into account variables given in Table 1, the objective was to provide a conceptual model with clearer influences and interactions in the soil and climate variables for each of the characteristics observed in the experiments. As a result, the conceptual model shown in Fig. 1 was constructed. This model was based on (as much explicit as) in tacit knowledge and it was built with the help of experts in plant physiology. It is possible to observe the plant, soil and climate interactions and identify the variables that should be collected for preparing the database.

Regarding the climate variables (Table 1), it was necessary to define the data granularity that was most suitable for manipulation (daily, weekly, every two weeks or monthly basis). Since the maize planting cycle ranges from 120 to 180 days, the data collected on a daily basis would generate a large volume of information and during certain periods with little variability. In interviews with experts in plant physiology and agro-meteorology, it was established that a daily variation of climate data would not affect plant growth, and that this data could be used in an aggregate manner (month-based values).

On the other hand, the development of the maize plant is divided into stages. Seven major stages were defined with the help of experts in plant physiology where the data regarding plant growth was collected (see Table 2).

The change of stage is defined by the evolution of the plant characteristics such as the number of leaves, silk appearance (-maize hair), among other data. Since this data was not observed during the experiments, the average amount of days in which each

Table 1
variables collected for each experiment of 2003/2004.

Variables–plant	Description
Cultivar	Name of cultivar
Cob weight (kg/ha)	Cob weight adjusted for kg/ha, with the corrected humidity degree for 13%, from the cob weighing harvested in the area
Grain weight (kg/ha)	Grain weight adjusted for kg/ha, with the corrected degree of humidity to 13%, from the weighing of grain all harvested ears in the field
Flowering (days)	Duration in days elapsed from appearing to female flowering (style-stigma emission) in 50% of the plant portion
Plant height (cm)	Plant height measured in cm from the soil surface to the insertion of the tassel with the flag leaf
Cob height (cm)	Cob height, measured in cm from the soil surface to the insertion of the upper cob
Lodged + broken (%)	Percentage of lodged and broken[a] plants
Final stand	Final stand determined by the number of plants per portion
Cob amount	Total number of cobs in the portion
Cob disease (%)	Percentage of sick cobs in the portion
Humidity (%)	Average humidity of the grain in the portion
Variables–soil	Description
Saturation per basis (V%)	The higher the *V* value the more fertile soil is. Regarding productivity, the base saturation of 50–60% is considered appropriate for most cereals and between 60% and 70% adequate for most leguminous
Texture	The soils are grouped into three classes of texture: Sandy, average and Clayey. Soils of sandy texture (light soil) contain sand levels higher than 70% and clay levels below 15%. Soils with average texture (medium soil) contain a certain balance between the proportions of sand, clay and silt. Clayey soils have a clay content above 35%
Available water	Fraction of the water present in the soil which is capable of being absorbed by plant roots
Elevation	It is the vertical distance from a point that separates it from the average sea level
Variables–climate	Description
Maximum temperature	It is the highest recorded temperature at a given point in space
Minimum temperature	It is the lowest temperature recorded at a particular point in space
Precipitation	Atmospheric precipitation in the form of drops of water (rain) measured by rain gauges and expressed through a value in mm
Wind	Movement of air, usually from an area of higher pressure to an area of lower pressure
Solar radiation	Radiation emitted by the sun in short wavelengths (between 0.15 and 4 mm), which correspond to about 99% of the solar radiation reaching the Earth. It is given in number of hours of sunshine during a day
Relative air humidity	The relationship between the amount of water vapor contained in the air and the maximum amount of vapor the air can hold under the same conditions of temperature and pressure. It is expressed as a percentage

[a] Lodged plants contain a layered angle of inclination equal to or less than 45° to the ground. Broken plants are plants at the harvest which stems have broken below the cob insertion.