Diagnosing soybean diseases from symptoms by means of evolutionary support vector machines

Ruxandra Stoean[1], Catalin Stoean[1], Mike Preuss[2] & Dan Dumitrescu[3]

[1] Department of Computer Science, University of Craiova, 13 Al. I. Cuza St., 200585 Craiova, Romania, e-mail: ruxandra.stoean@inf.ucv.ro; catalin.stoean@inf.ucv.ro

[2] Department of Computer Science, University of Dortmund, 14 Otto-Hahn-St, 44221, Dortmund, Germany, e-mail: mike.preuss@cs.uni-dortmund.de

[3] Department of Computer Science, Babes-Bolyai University of Cluj-Napoca, 1 Mihail Kogalniceanu St., 400084 Cluj-Napoca, Romania, e-mail: ddumitr@cs.ubbcluj.ro

Received: May 04, 2006 ▷ Accepted: July 29, 2006

Abstract. Soybean (*Glycine max*) diseases are difficult to diagnose due to the similarity of symptoms and to deviations between indicators of a given disease as a result of differences in local conditions or over time change. The paper proposes a novel computational technique, *evolutionary support vector machines*, that can differentiate four types of diseases found in soybean. The diagnosis is based on comparing symptoms displayed by the considered plants and those of previously predicted cases. The proposed, highly accurate method provides a way of validating expert decision-making and broadening of knowledge, and will assist in the management of these diseases.

Key words: canker, disease diagnosis from symptoms, evolutionary support vector machines, rot, *Phytophtora* rot, *Rhizoctonia* root rot, soybean

Introduction

Soybean diseases are of major concern as it has been estimated that they reduce the world's annual soybean production by 11% per year, an equivalent of approximately 15 million tons (Ryley 2003). The most serious diseases include canker (*Diaporthe phaseolorum* var. *caulivora*) and rot (*Macrophomina phaseolina*). Other important diseases, such as *Phytophthora* rot caused by *Phytophthora sojae*, affect yield primarily through the reduction of soybean populations; in addition, *Phytophtora* produces damage to the protein content of the seed.

Computational techniques that could identify and differentiate each type of soybean disease, based on comparison of symptoms displayed by the considered plants and those of previously predicted cases, would give a helping hand in controlling the diseases.

The paper proposes a method based on the state-of-the-art computational heuristics related both to machine learning and optimization. Evolutionary support vector machines constitute a hybrid paradigm between a successful learning technique represented by support vector machines and powerful optimization methods, named evolutionary algorithms.

Tests to validate the efficiency of the concept in diagnosing a specific type of disease in soybean plants were conducted on the soybean (small) database from the University of California at Irvine (UCI) Repository of Machine Learning Databases[1]. The data set contains information about plants suffering from 4 types of soybean diseases, i.e. the three aforementioned, together with *Rhizoctonia* root rot.

[1] http://www.ics.uci.edu/~mlearn/MLSummary.html

Material and methods

The soybean collection contains 47 plant instances, each recording 35 attributes plus one attribute reflecting the corresponding disease. There are no missing values. The first 35 attributes enclose information pertaining to the description of the environmental attributes (e.g. normality of air temperature, precipitation), the plant's global attributes (e.g. seed treatment, plant height) and the local attributes (i.e. condition of leaves, stem, fruits – pods, seed and roots). A list of these attributes is given in Table 1 (Michalski & Chilausky 1980). However, an indicator of the leaf spot shape is missing from the UCI data set.

The last attribute of every record shows the disease of the current soybean plant (the plant that a diagnosis is being made on). There are four diseases considered in this paper, i.e. canker, rot, *Phytophtora* rot, and *Rhizoctonia* root rot. Indicators for each disease are presented below (Colyer 2002; Soybean Plant Health 2006).

One highly destructive disease, canker (*D. phaseolorum* var. *caulivora*), kills soybean plants from flowering to maturity. The fungus lives on diseased stems and seeds over winter. Canker has an increased incidence during seasons with an elevated level of humidity. First symptoms appear after flowering and are characterized by small reddish-brown lesions at the

Table 1. Description of attributes for each plant in the considered soybean collection.

Category	Attribute	Possible values
Environmental attributes	Time of occurrence	April, May, June, July, August, September, October
	Plant stand	Normal, less than normal
	Precipitation	Less than normal, normal, above normal
	Temperature	Less than normal, normal, above normal
	Occurrence of hail	Yes, no
	Number years crop repeated	Different last year, same last year, same last 2 years, …, same last 7 years
	Damaged area	Scattered, groups of plants in low areas, groups of plants in upland areas, whole fields
Plant global attributes	Severity	Minor, potentially severe, severe
	Seed treatment	None, fungicide, other
	Seed germination	90-100%, 80-89%, less than 80%
	Plant height	Normal, abnormal
Plant local attributes	Condition of leaves	Normal, abnormal
	Leaf spots – halos	Absent, with yellow hallows, without yellow hallows
	Leaf spots – margin	With water soaked margin, without water soaked margin, does not apply
	Leaf spot size	Less than 1/8″, greater than 1/8″, does not apply
	Leaf shredding or shot holing	Absent, present
	Leaf malformation	Absent, present
	Leaf mildew growth	Absent, upper surface, lower surface
	Condition of stem	Normal, abnormal
	Presence of lodging	Yes, no
	Stem cankers	Absent, below soil line, at or slightly above soil line, above second node
	Canker lesion color	Brown, dark-brown or black, tan
	Fungal fruiting body on stem	Absent, present
	External decay	Absent, firm and dry, watery
	Mycelium on stem	Absent, present
	Internal discoloration	None, brown, black
	Sclerotia – internal or external	Absent, present
	Condition of fruits - pods	Normal, diseased, few present, does not apply
	Fruit spots	Absent, colored, brown spots with black specks, distorted, does not apply
	Condition of seed	Normal, abnormal
	Mold growth	Absent, present
	Seed discoloration	Absent, present
	Seed size	Normal, less than normal
	Seed shrivelling	Absent, present
	Condition of roots	Normal, rotted, galls and cysts

level of branches and petioles. These lesions extend, become dark-brown or black, sunken and surround the stem. Occasionally, the pathogen may be active within the stem but cause no visible lesion. The leaves display a yellow color between the veins that turns necrotic. On dead plants, the leaves are shrivelled but remain attached to the stem.

The symptoms of charcoal rot (*Macrophomina phaseolina*) can be noticed primarily on roots, and at the stem base, but can also appear higher above. The taproot and lower stem have a grayish discoloration. Ultimately, an infected plant loses vigour and dies. Sclerotia, i.e. many small, black specks, below the bark of root and lower stem are a diagnostic symptom of charcoal rot. The name of the disease comes from the appearance of sclerotia on the plant that resembles a sprinkling of charcoal. The disease can reduce plant height, root volume and weight by more than 50 %, as well as seed number and weight. Seeds become smaller in number and weight. Plants may also show a premature yellowing on the top leaves, or leaf drop (Colyer 2002) (Soybean Plant Health 2006).

Phytophthora root and stem rot (*Ph. sojae*) appears where soils are poorly drained, but can occur in normally well-drained fields that are saturated for 7–14 days due to excessive precipitation or irrigation (Nicholls 2004). The fungus survives as thick-walled spores (oospores) in plant debris and in soil. Symptoms appear at any stage of growth. Not all infected plants die but are less productive than the healthy ones. Symptoms include reduced plant stands, dark lesions on stems and roots of seedlings often leading to death, and on older plants, yellowing of leaves which bend down but remain attached to the stem, loss of vigour and a diagnostic dark-brown lesion on the lower stem. *Phytophthora sojae* causes loss of more than US$ 1,000,000,000 per year to soybean growers.

Rhizoctonia root rot (*Rh. solani*) can be noticed on dead plants in the beginning of summer. Infected seedlings have a red-brown lesion on the hypocotyl and often die, older infected plants are stunted with dark-reddish lesions at the stem base and on roots, lateral roots are pruned, and infected plants are often in circular patches (Yang 1996; Bradley & al. 2001; Soybean Disease Research 2006).

Some features of the symptoms associated with these four soybean diseases are similar (Kersten & al. 1999). A computational means of establishing the specific disease based on given symptoms and previous diagnosis of other such plants could prove to be very useful in diagnosing the disease and commencing its management.

A new supervised learning and optimization technique, called evolutionary support vector machines (ESVMs) (Stoean & al. in press), has the potential to solve the soybean disease diagnosis problem mentioned above.

As all supervised learning machines, ESVMs act in two stages. In the training stage, the black-box containing the correspondence between every data sample (plant symptoms), namely $x_i \in R^n$, and given class (disease), y_i, $i = 1,2,...,m$, is internally opened and learnt. In the test step, prediction of the disease for previously undiagnosed plants is performed, according to what has been learnt.

ESVMs derive from the well-known paradigm of support vector machines (SVMs), from which they have inherited the formulation of the learning problem. Consequently, ESVMs solve any multi-class classification task (as is the case in the soybean problem) by building pairs of two-class subproblems, solving them and then applying a voting system.

Additionally, as two-class classifiers, i.e. $y_i \in \{-1,1\}$, ESVMs regard the learning process in a geometrical fashion, i.e. they assume the existence of a linear separating surface between classes, where data samples lie either on its positive or negative side, according to their class. Subsequently, the coefficients of this decision *hyperplane* have to be found, i.e. its orientation (denoted by the *n*-dimensional *w* vector) and location (denoted by *b*) and search has to be based on the attributes of training data and corresponding classes (Fig. 1a). Nevertheless, since data are usually non-linearly separable, ESVMs also allow for some errors; indicators for errors are namely $\xi_i, i = 1,2,...,m$, which exceed unity when signaling an error (Fig. 1b). Furthermore, an extension is also available to handle this situation in a nonlinear and natural fashion: a procedure that maps the data into a higher dimensional space, where they can be linearly separated, is conducted; this leads to a nonlinear separating surface in the initial space (Fig. 2).

Since ESVMs impose that the separating surface to be found has to generalize well, i.e. the accuracy on the particular training set and the capacity of the machine to learn any training set without error are well balanced, all the concepts above can be expressed as the following optimization problem, which is the core of ESVMs:

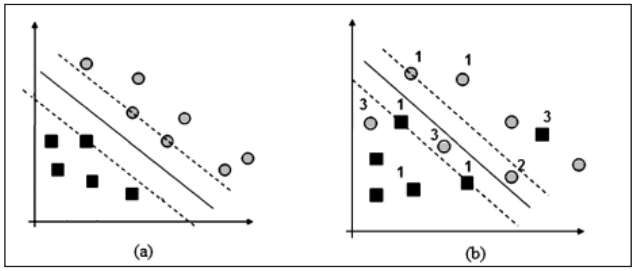

Fig. 1. A linear separating surface between classes, i.e. the middle line; the two dotted lines are the supporting surfaces for each of the two classes (a). Separating and supporting surfaces between classes and indicators for errors (deviations from the ideal condition of separation); label 1 corresponds to $\xi_i = 0$, which means data vector correctly placed; label 2 matches $\xi_i < 1$, which is still correctly placed (in the margin); and label 3 denoted, $\xi_i > 1$ which means error of classification (b).

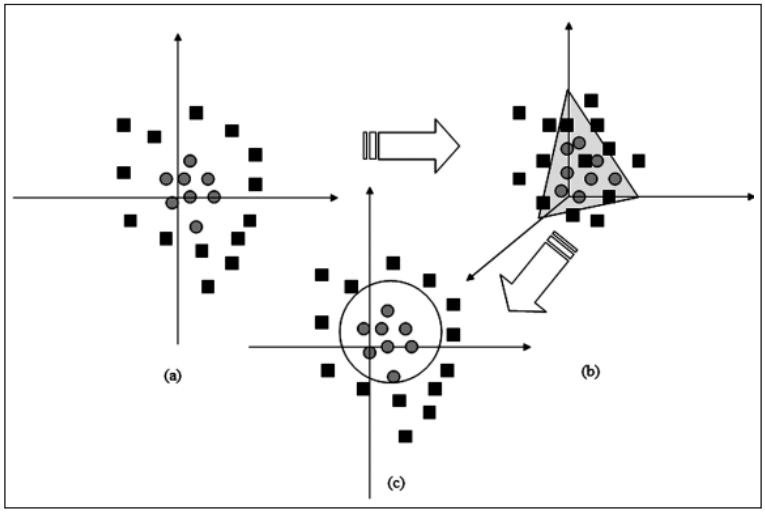

Fig. 2. A situation of nonlinear separable data (a). The data are mapped into a higher dimensional space where they are linearly separated (b). This corresponds to a nonlinear (radial) separating surface for the initial data (c).

Given the training set $\{(x_i, y_i)\}_{i=1,2,...,m}$, find the *optimum* values for *w* and *b* so as to *minimize* the objective function

$$\frac{1}{2}K(w,w) + C\sum_{i=1}^{m}\xi_i \ , \ C > 0$$

subject to constraints $\begin{cases} y_i(K(w,x_i) - b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$,

$i = 1,2,...,m$, where *K* is called a *kernel* and can either be polynomial, $K(x,y) = \langle x, y \rangle^p$ (where *p*=1 corresponds to the linear case), or radial, $K(x,y) = e^{\frac{\|x-y\|^2}{\sigma}}$. The kernel is the function that is able to map the initial nonlinearly separable data into the higher dimensional space, where samples are linear. The above formulation of the optimization problem expresses the learning problem that has to be solved, in order to determine the hyperplane that linearly separates data in the higher space and which can be reversely, through the means of the same kernel, interpreted into the corresponding nonlinear function of the initial space.

The objective function in the formulation above corresponds to the requirement that the machine can also learn other training sets, while the constraints formally translate the condition that the hyperplane separates well among data of the two classes. Indicators for errors are also present in the ESVM formulation, in order to allow for some classification errors as long as they are kept to a minimum; hence, the inclusion of the sum of indicators in the objective function with the presence of a variable that assigns penalty for the errors (denoted by *C*).

Accordingly, either a radial separating surface (Fig. 2), or a polynomial (odd or even) one (Fig. 3) is obtained, subsequent to the solving of the optimization problem.

Now, parting from the mother paradigm of SVMs (which employs a complex method of Lagrange multipliers), ESVMs reach the solution to the optimization problem above through a simpler and more direct way, i.e. by applying a standard evolutionary algorithm (EA) (Eiben

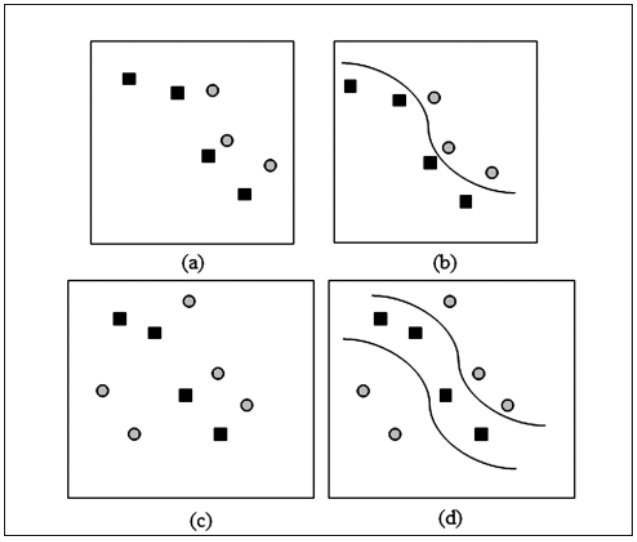

Fig. 3. Other situations of nonlinear separable data (a), (c). This corresponds to a nonlinear separating surface, i.e. odd polynomial in the first case (b) and even polynomial in the second situation (d).

& Smith 2003). The idea behind an EA is based on principles of the theory of evolution and laws of genetics. Having a fitness function to be optimized, a set of randomly generated candidate solutions (or individuals) are created in the domain of the function and are evaluated – fitter individuals (or solutions) are considered those with better values for the fitness function. The evolutionary process starts when, based on the fitness values, some of them are selected to be the parents of the population in the next generation by applying recombination and/or mutation to them. Recombination takes place between two or more individuals and one or more descendants (or offspring) are obtained; descendants borrow particularities from each of the parents. When mutation is applied to a candidate solution, the result is a new candidate that is usually only slightly different from its parent. After applying these variation operators, mutation and recombination, a set of new individuals is obtained that will fight for survival with the old ones for a place in the next generation; the candidate solutions that are fitter are advantaged in this competition. The evolutionary process resumes and usually stops after a predefined computational effort limit is reached.

In ESVMs, an individual is represented by the coefficients of the hyperplane, i.e. *w* and *b*, together with indicators for errors of separation, i.e. $\xi_i, i = 1,2,...,m$. The fitness function is defined as to comprise the objective criterion and penalize the individuals that violate the constraints through the penalty function *t*:

$$f((w_1, w_2, ..., w_n, b)) = w_1^2 + ... + w_n^2 + \sum_{i=1}^{m} [t(y_i(\langle w, x_i \rangle - b) - 1)]^2$$

where $t: R \to R, t(a) = \begin{cases} a, a < 0 \\ 0, otherwise \end{cases}$, and one is led to *minimize (f(c), c)*.

In the end of the algorithm, the coefficients of the separating surface, i.e. *w* and *b*, are obtained and the class for a new data sample is appointed, following the sign of the resulting hyperplane.

In order to apply ESVMs to the soybean diagnosis problem discussed above, which is 4-class, a number of six two-class ESVMs classifiers are built, every time bringing one class against another. When coefficients of every separating surface are found, a voting system is applied in order to decide the disease for a new soybean plant.

Results

Before validating the new approach as an appropriate means of diagnosing soybean diseases based on symptoms, the following settings were performed. For 30 times, the collection was randomly split into 75 % training – 25 % test cases.

Parameters of the support vector machine and EA were manually chosen at first and subsequently adapted to the problem by means of a state-of-the-art parameter tuning method called sequential parameter optimization (SPO) (Bartz-Beielstein 2006). Starting from a Latin hypercube sample (LHS), this method employs a nonlinear regression model on the parameters of an optimization algorithm, supported by kriging for error estimation. Parameter configurations that promise high expected improvement are then tested in a sequential manner, concurrently incorporating newly available information into the model to increase its accuracy. Moreover, SPO is also an integrated approach for applying proper statistical techniques, as hypothesis testing. For the classification problem treated in this work, it enabled finding suitable parameters for the embedded EA, resulting in a near-optimal ESVM classification result.

Accuracy is computed on the test set and its value is given by the percent of cases that were correctly classified by ESVMs. Obtained results are depicted in Table 2 and prove the suitability and success of the method in diagnosing a type of soybean disease in new plants, based on symptoms and disease of earlier diagnosed ones.

Table 2. Obtained test accuracy in 30 runs for forecasting the disease for new, undiagnosed soybean plants through ESVMs.

Parameter tuning	Descriptors	Test accuracy
Manual	Average	99.02 %
	Worst	94.11 %
	Best	100 %
	Standard deviation	2.23 %
SPO	Average	99.80 %
	Worst	94.11%
	Best	100 %
	Standard deviation	1.05 %

Comparison to other previously applied techniques for the soybean task was also conducted. Unfortunately, a direct comparison to standard SVMs could not be performed, as no results for this data set were found in

literature. Accuracy for the soybean diagnosis problem was reported by Bailey & al. (2003), where classification with constraint emerging patterns and the Naïve Bayes method were conducted. The former method provided an accuracy of 95.50 % (reaching 100 % when a pairwise classification strategy was employed), while the latter resulted in 98 % accuracy.

Discussion

The aim of this paper was to demonstrate the suitability and ability of a new computational learning technique to identify a disease in soybean plants based on symptoms and knowledge of previous cases. Data came from the UCI repository of machine learning databases and used 35 indicators to predict the class for 4 types of soybean diseases: rot, *Phytophtora* rot, *Rhizoctonia* root rot, and canker.

The novel ESVMs were conceptually envisaged as a simple, i.e. no complex mathematics, well performing tool for difficult classification problems. As a consequence, ESVMs inherited the high classification power of standard SVMs, which is due to the natural separation model they implement. On the other hand, ESVMs enhanced the performance of their parent through the use of the powerful optimizers that EAs represent. As a result, the new technique brings a simpler and more straightforward alternative to the solving of the inherent optimization problem. Moreover, the coefficients of the separating surface are acquired in a direct manner, opposite to standard SVMs, where most of the time they cannot be obtained at all (more mathematical artifices are performed in order to label new data).

The obtained results prove the suitability of proposed technique in checking the consistency of decision-making when labeling soybean plants as affected by a certain type of disease. The method can be successfully applied to handle other categories of soybean diseases.

In conclusion, ESVMs are both a simpler alternative to other classifiers and also provide very good and accurate results for the given soybean task. The methodology can nevertheless be broadened successfully to handle other prediction tasks from the phytological domain.

References

Bailey, J., Manoukian, T. & Ramamohanarao, K. 2003. Classification using constrained emerging patterns. – In: **Dong, G., Tang, C. & Wang, W.** (eds), Proc. of WAIM 2003, Lecture Notes in Computer Science Series. Pp. 226-237. Springer Verlag, Berlin & New York.

Bartz-Beielstein, T. 2006. Experimental research in evolutionary computation – the new experimentalism. – Natural Computing Series, Springer Verlag, Berlin & New York.

Bradley, C.A., Hartman, G.L., Nelson, R.L., Mueller, D.S. & Pedersen, W.L. 2001. Response of ancestral soybean lines and commercial cultivars to Rhizoctonia root and hypocotyl rot. – Plant Dis., **85**: 1091-1095.

Colyer, P.D. (ed.) 2002. Soybean Disease Atlas. 2nd ed. – Southern Soybean Disease Workers, USA; http://cipm.ncsu.edu/ent/SSDW/soyatlas.htm

Eiben, A.E. & Smith, J.E. 2003. Introduction to Evolutionary Computing. Springer Verlag, Berlin & New York.

Kersten, G.E., Mikolajuk, Z. & Yeh, A. (eds). 1999. Decision Support Systems for Sustainable Development: A Resource Book of Methods and Applications. Springer Verlag, Berlin, New York.

Michalski, R.S. & Chilausky, R.L. 1980. Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. – Int. J. Policy Analysis Inform. Systems, 4(2): 125-160, http://www.mli.gmu.edu/papers/79-80/80-2.pdf

Nicholls, H. 2004. Stopping the rot. – PloS Biology, **2**(7): e213; http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0020213

Ryley, M. 2003. Effects of some diseases on the quality of culinary soybean seed. – Proc. 12th Australian Soybean Conf., Toowoomba. Northern Australian Soybean Industry Association, Toowomba; http://www.australianoilseeds.com/__data/page/269/Malcolm_Ryley-Effects_of_some_diseases_on_the_quality_of_culinary_soybean_seed.pdf

Soybean Disease Research. 2006. http://www.soydiseases.uiuc.edu/index.cfm

Soybean Plant Health. 2006. University of Wisconsin-Madison, Departments of Agronomy, Entomology and Plant Pathology; http://www.plantpath.wisc.edu/soyhealth/index.htm

Stoean, R., Dumitrescu, D. & Stoean, C. In press. Nonlinear evolutionary support vector machines. Application to classification. – In: **Frentiu, M.** (ed), Stud. Univ. Babes-Bolyai, Ser. Inform., Cluj-Napoca.

Yang X.B. 1996. Soybean root rot diseases are here. Iowa State Univ., Integrated Crop Management Newsletter, Plant Diseases; http://www.ipm.iastate.edu/ipm/icm/1996/7-15-1996/soyrootrothere.html