Data Mining Techniques in Agricultural and Environmental Sciences

Altannar Chinchuluun, Department of Industrial and Systems Engineering, , University of Florida, USA

Petros Xanthopoulos, Department of Industrial and Systems Engineering, , University of Florida, USA

Vera Tomaino, Department of Industrial and Systems Engineering, , University of Florida, USA; Department of Experimental and Clinical Medicine, , University Magna Græcia of Catanzaro, Italy

P.M. Pardalos, Department of Industrial and Systems Engineering, , University of Florida, USA

ABSTRACT

Data mining techniques are largely used in different sectors of the economy and they increasingly are playing an important role in agriculture and environment-related areas. This paper aims to show our vision on the importance of knowing and efficiently using data mining and machine learning-related techniques for knowledge discovery in the field of agriculture and environment. Efforts for searching hidden patterns in data are not a recent phenomenon. History shows that extensive observations on data have helped discover empirical laws in different fields of research. Therefore, it is important to provide researchers in agriculture and environmental-related areas with the most advanced knowledge discovery techniques. Data mining is the process of extracting important and useful information from large sets of data. This information can be converted into useful knowledge that could help to better understand the problem in study and to better predict future developments. The paper presents the state of the art in data mining and knowledge discovery techniques and provides discussions for future directions.

Keywords: Agriculture; Artificial Neural Networks; Data Mining; k Nearest Neighbor; k-Means; Optimization; Support Vector Machines

Keywords: Agricultural and Environmental Sciences, Data Mining Techniques

INTRODUCTION

The problem of searching for patterns in data is a fundamental one and has a long and successful history. There are many examples in different research areas that extensive observations of data has led to discovering empirical laws. As an example, the attentiv astronomical observations undertaken by several astronoms allowed

DOI: 10.4018/jaeis.2010101302

Kepler to discover the laws of planetary motion (Bishop, 2006).

Over the years, several techniques have been developed to discover hidden patterns in data and these efforts led to the creation of a rigourous discipline known as data mining or knowledge discovery. Data mining is the process of finding useful patterns or correlations amongst data. These patterns, associations, or relationships between data can provide information about the problem in study and information can then be transformed into knowledge. The idea of using information hidden into relationships amongst data inspired researchers to apply these techniques for predicting future trends (Mucherino, Papajorgi, & Pardalos, 2009). Data mining techniques are developed from mainly three areas: statistics, artificial intelligence and machine learning. Although the roots of data mining may seem different, but essentially they aim the same target: discover a relationship that more or less maps measurements in one part of a data set to measurements in another, linked part of the data set (Pyle, 2003).

Regardless of the method used, the goal of data mining techniques is to split data in different categories, each of them representing some feature of interest the data may have (Mucherino et al., 2009). Thus, fundamental for the success of a data mining technique is the ability to group available data in disjoint categories, where each category contains data with similar properties. The similarity between data is usually measured using a distance function; similar data should belong to the same group or cluster. Therefore, the success of a data mining technique depends on the adequate definition of a suitable distance between data samples.

As the similarity between data samples is measured using a distance function, often it occurs that this distance needs to be optimal. Thus, many data mining techniques led to the formulation of a global optimization problem (Mucherino et al., 2009).

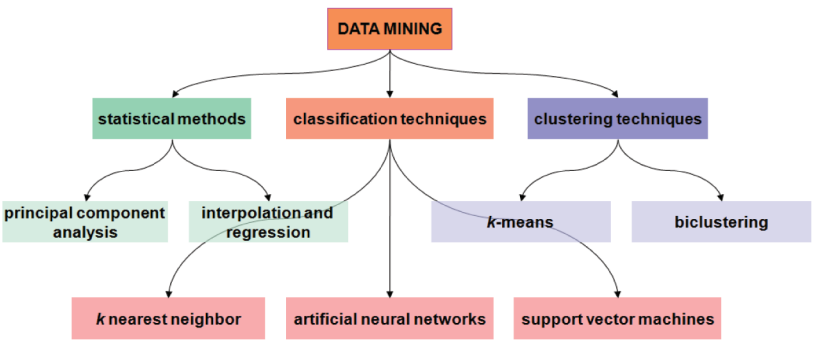
Data mining techniques can be grouped in three categories as shown in Figure 1.

Statistical Methods

Statistical methods such as Principal Component Analysis (PCA) and regression techniques are commonly used as simple methods for finding patterns in sets of data. PCA is a useful statistical technique that has found application in fields such as image compression, and is a common technique for finding patterns in big data sets. PCA helps identifying patterns in data, and expressing the data in such a way as to emphasize their similarities and differences. Since patterns in large data sets can be hard to find, because graphical representation of the data is not available, PCA is a powerful tool for analyzing data.

The main advantage of PCA is that once patterns in the data are identified data can be represented as components ordered by their

Figure 1. A schematic representation of the classification of the data mining techniques



relevance and it is possible then to discard components of low level of relevance without loss of important information and thus, reducing the complexity of the problem. In many cases, dimension reduction makes it possible to represent data graphically that enormously facilitates the understanding of discovered patterns.

Classification Techniques

Classification techniques use the well-known principle of Cicero *pares cum paribus facillime congregantur* (*birds of a feather flock together* or literally *equals with equals easily associate*). These techniques try to classify an unknown sample based on the known classification of its neighbors (Mucherino et al., 2009). Let us suppose that a set of samples with known classification is available, that is referred to as a training set. Each unknown sample should be classified considering its surrounding samples. Therefore, if the classification of a sample is unknown, then it could be predicted by considering the classification of its nearest neighbor samples. The main idea behind the classifications method is that given an unknown sample and a training set, all the distances between the unknown sample and all samples in the training set can be computed. The distance with the smallest value corresponds to the sample (or samples) in the training set closest to the unknown sample. Therefore, the unknown sample may be classified based on the classification of this nearest neighbor.

Clustering Techniques

Clustering techniques are used for partitioning a given set of data samples in clusters and there is no knowledge a priori about these data (Arulselvan, Baourakis, Boginski, Korchina, & Pardalos, 2008, Mucherino et al., 2009). The main idea behind the clustering algorithms is partitioning a set the data into a number *of* disjoint clusters by looking for inherent patterns in the set.

Clustering techniques are divided in *hierarchical* and *partitioning*. The hierarchical clustering approach builds a tree of clusters. The root of this tree can be a cluster containing all the data. Then, branch by branch, the initial big cluster is split in sub-clusters, until a partition having the desired number of clusters is reached. In this case, the hierarchical clustering is referred to as *divisive*. Furthermore, the root of the tree can also consist of a set of clusters, in which each cluster contains one and only one sample. Then, branch by branch, these clusters are merged together to form bigger clusters, until the desired number of clusters is obtained. In this case, the hierarchical clustering is referred to as *agglomerative*.

The paper is organized as follows. We review the main optimization based data mining techniques, including *k*-means clustering techniqe and support vector machine classifications, in the next section. The applications of these data mining techniques in agricultural engineering will be presented and future directions of this field are discussed.

Data Mining Algorithms

Data mining is the process of analyzing data using tools such as clustering, classification, feature selection and outlier's detection. Clustering techniques partition a given set of data into groups of similar samples according to some similarity criteria. Classification techniques determine classes of the test samples using known classification of a training data set. Feature selection techniques select a subset of features responsible for creating the condition corresponding to any class. Clustering is generally an initial step of data mining and it groups data into similar samples which can be used as a starting point of other techniques. Data clustering can be divided into two parts as hierarchical and partitional clustering (Jain, Murty, & Flynn, 1999). Single link and complete link are examples of hierarchical clustering while partitional clustering includes squared error algorithms (*k*-means), graph theoretic, mixture resolving (expectation maximization), mode seeking, and so forth. Bayesian classifier is a traditional statistical classification algorithm

based on Bayes' theory. Bayesian classifiers use combination of conditional probability and posterior probabilities for classifying any information. Further details about Bayesian classifiers and applications can be found in Duda and Hart (1973), Marchant and Onyango (2003), Pernkopf (2005) and Yager (2006). Some other clustering techniques such as fuzzy clustering, artificial neural networks, nearest neighbor clustering and evolutionary approach based clustering are also becoming very popular tools for researchers. We discuss some of these clustering techniques in this section.

The k-means Algorithm

The *k*-means (MacQueen, 1967) method is one of the most popular unsupervised learning or clustering methods which have been applied in a variety of fields including pattern recognition, information retrieval, document extraction and microbiology analysis, and so forth. The method is called the *k*-means because it represents each of *k* number of clusters C_j (*j*= 1,2,…,*k*) by the mean (or the weighted average) of its points. The goal of this method is to classify a given data set through a certain number of clusters such that some metric relative to the centroids (or centers) of the clusters is minimized. We can define our problem mathematically as follows.

Suppose that we are given a set *X* of a finite number of points in *d*-dimensional Euclidean space R^r, that is, $X = (x^1, x^2, ..., x^n)$ where $x^i \in R^d$, *i*=1,2,…,*n*.

We aim at finding a partition C_j, *j*=1,2,…,*k*:

$$X = \bigcup_{j=1}^{k} C_j, \quad C_j \bigcap Cl = 0,$$

for all $j \neq l$, of *X* which minimizes the squared error function

$$f(C_1, C_2, ..., C_k) = \sum_{j=1}^{k} \sum_{x^i \in C_j} \| c^j - x^i \|^2,$$

where $\| \cdot \|$ denotes the Euclidean norm, C^j is the center of the cluster C_j

$$c^j = \frac{1}{| C_j |} \sum_{x^i \in C_j} x^i, \quad j = 1,2,...,k \qquad (1)$$

Algorithm k-Means

Step 1. Initialize the centroids

$$C\frac{j}{0},$$

j=1,2,...,k . Set *q*=0 (where *q* is the iteration counter).

Step 2. Assign each point x^i (*i*=1,2,…,*n*) to the cluster that has the closest centroid (or the center of the cluster) $c^j (j \in \{1,2,..k\}$, that is $j = ar\,gmin_{1 \le l \le k} \le \| x^i - c_q^l \|^2$.

Step 3. When all points have been assigned, for *j*=1,2,…,*k*, calculate the new position c_{q+1}^j of the centroid *j*.

Step 4. If $c_q^j = c_{q+1}^j$ for all *j*=1,2,…,*k*, then stop, otherwise set *q*=*q*+1 and go to Step 2.

The *k*-means algorithms is easy to implement and its time complexity is order of *n* (*O(n)*), where *n* is number of patterns (Jain et al., 1999). However, it finds one of the many local solutions that depend on the initial starting points. To find a better solution, we can run the algorithm several times and choose the best one as the optimal solution. Unfortunately, repetition with different random selections (Duda & Hart, 1973) appears to be not a very efficient method. For this purpose, Bradley and Fayyad (1998) presented a procedure for computing a refined starting condition from a given initial one that is based on an efficient sampling technique for estimating the modes of a distribution, and their experiments presented that refined initial starting points indeed lead to improved solutions. Yager and Fillev (1994) developed the *mountain method* which is a simple and effective approach for approximate estimation of the cluster centers on the basis of the concept of a mountain function. It can be useful for obtaining the initial values of the clusters that are

required by more complex cluster algorithms. Another drawback of the algorithm is that there are no efficient methods for defining the initial number of partitions. Many alternative methods to improve *k*-means were published in literature. Krishna and Murty (1999) proposed a new Genetic *k*-means algorithm (GKA) for global search and faster convergence. Zhang, Xiong, Mao, and Ou (2006) proposed parallel *k*-means algorithm for higher efficiency. Many algorithms similar to *k*-means have been appeared in the literature (Duda, Hart, & Stork, 2001; MacQueen, 1967).

Fuzzy c-means (FCM) Clustering

In the *k*-means algorithm, each sample can be assigned to only one cluster. Fuzzy clustering relaxes this restriction and an object can belong to several clusters at the same time but with certain degrees of memberships. The most known fuzzy clustering method is the fuzzy *c*-means method (FCM), introduced by Dunn (1974) and later generalized by Bezdek (1981). FCM partitions a data set $X = (x_1, x_2, ..., x_n) \subset R^p$ of *p* features, into *c* fuzzy subsets where $u_{i,k}$ is the membership of x_k in class *i* (*i*=1,2,…,*c*). These classes are identified by their cluster centers v_i, (*i*=1,…,*c*). The objective of FCM is to find an optimal fuzzy *c* partition minimizing the objective function,

$$Jm(U,V:X) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} \parallel x_k - v_i \parallel^2, \quad (2)$$

where the value of fuzzy partition matrix *U* is constrained in the range [0,1] such that

$$\sum_{i=1}^{c} u_{ij} = 1, \quad k = 1,2,...,n \quad (3)$$

and

$$\sum_{k=1}^{n} u_{ik} \le n, \quad i = 1,2,...,c \quad (4)$$

Here, $m \in [1,\infty)$ is an exponential weighting function that controls the fuzziness of the membership values, $\|\cdot\|$ is the Euclidean norm and $V = (v_1, v_2, ..., v_c)$ is a matrix of unknown cluster centers $v_i \in R^p$ (*i*=1,…,*c*). Fuzzy *c*-means algorithm to minimize (2):

Step 1. Choose appropriate values for *m* and *c,* and a small positive number $\in$. Initialize randomly a fuzzy partition matrix U^0 and set iteration number *t*=0.

Step 2. For given membership values $u_{ik}^{(t)}$, calculate the cluster centers $v_i^{(t)}$ (*i*=1,2,…,*c*) as

$$v_i^{(t)} = \frac{\sum_{k=1}^{n} (u_{ik}^{(t)})^m x_k}{\sum_{k=1}^{n} (u_{ik}^{(t)})^m} \quad (5)$$

Step 3. Given a new cluster center from *Step 2*, update membership values $u_{ik}^{(t+1)}$ using

$$u_{ik}^{(t+1)} = \left[\sum_{j=1}^{c} \left(\frac{\| x_k - v_i^{(t)} \|^2}{\| \sum_{k=1}^{n} \left(u_{ik}^{(t)} \right)^m \|} \right) \right] \quad (6)$$

Step 4. Repeat *Step 2* and *3* until $|U-(t+1)-U^{(t)}| < \in$ or a pre-defined number of iterations is reached.

Methods discussed in previous sections are crisp/hard partitioning methods, which allow to partition data into a specified number of mutually exclusive datasets only, while fuzzy methods are soft partitioning methods where an object can belong to one or more data sets/partitions. Similarly to the crisp/hard partitioning methods, selection of initial matrix of centers plays an important role in convergence of FCM. Many times FCM does not guarantee the global optimal solutions due to randomized initialization of cluster centers and matrix U. Moreover, FCM solutions are also sensitive to noise and outliers. Hathway, Bezdek, and Hu (2000) have proposed a modified FCM using

1 norm distance to increase robustness against outliers. Hung and Yang's *psFCM* algorithm (Hung & Yang, 2001) finds the actual clusters' centers and refines initial value of FCM. This technique reduces the computational time by a large amount. Many other improved FCMs and their applications can also be found in (Hammah & Curran, 2001; Hathaway & Bezdek, 2001; Karayiannis, 1997; Zhang et al., 2006).

k-Nearest Neighbor Classification

The *k*-nearest neighborhood (Mitchell, 1997) method is widely adopted due to its efficiency. The key idea of the algorithm is to classify a new sample in the most frequent class of its closest neighbors in the training set. This is a majority voting formula on the class labels of its neighbors. A Euclidean distance measure is used to calculate how close each member of the training set is to the target data that is being examined. The *k*-nearest neighbor classification algorithm can be divided into two phases:

Training Phase

- Define a training set $S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i = (x_i^1, x_i^2, ..., x_i^d)$ is a *d*-dimensional feature vector of real numbers, for all *i*=1,…,*n*.
- Define class labels y_i corresponding to each x_i for all *i*, $y_i \in C$ where $C = (1, 2, ..., N)$, N_c is the number of different classes.
- Task: determine y_{new} for C_{new}

Testing Phase

- Find the closest point x_j to x_{new} with respect to Euclidean distance

$$\sqrt{(x_j^1 - x_{new}^1)^2 + \cdots + (x_j^d - x_{new}^d)^d}$$

- Classify by $y_{new} = y_j$

A serious drawback of this *k*-nearest neighbor technique is the computational complexity in searching the *k* nearest neighbors among those available training samples. Kuncheva (1997) claims to achieve better computational efficiency and higher classification accuracy by using genetic algorithms as editing techniques. Bermejo and Cabestany (2000) proposed a KNN algorithm with local Parzen window estimate to improve approximation quality. They also suggested an adaptive learning algorithm to allow fewer data points to be used in a training data set. Many other techniques have been proposed to reduce computational burden of *k*-nearest neighbor algorithms in Hwang and Wen (1998) and Pan, Qiao, and Sun (2004).

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) are efficient classification tools that have been applied to several applications including extracting regularities in data and classifying events in finance, marketing, Internet and biomedicine. The training process uses available examples to produce a model and to classify new events based on the extracted model. Neural networks are built from simple units, called neurons or cells by analogy with the way human brain works. Neurons are linked with each other by a set of weighted connections. The information to be analyzed is fed to the neurons of the input layer and then propagated to the neurons of the hidden layers (if any) for further processing. The result of this processing is then propagated to the next hidden layer and the process is continued until the output layer is reached. Each unit receives some information from other units and processes this information, which will be converted into the output of the unit. There are

no specific methods of choosing the network parameters such as number of hidden layers and type of activation function. Generally one input and one output nodes are chosen for each data class.

The training set is known a priori and is used to fine-tune the network for future similar records. While in training phase, known data containing inputs and corresponding outputs are fed to the network, and the network learns to infer the relationship between these two.

The process of classification by ANN can be broadly defined as follows:

- Run a sample from the training set, by giving its attribute values as input.
- The summation of weights and activation functions are applied at each node of hidden and output layers, until an output is generated (feed-forward process).
- Compare output with the expected output from training set.
- If output does not match, go back layer to layer and modify arc weights and biases of nodes (back-propagation process).
- Run the next sample and process the same.
- Eventually the weights will converge and process stops.

Feed forward topology is widely used in multilayer perceptions networks. Feed forward network provides a general framework for representing non-linear functional mappings between a set of input variables and a set of output variables (Bishop, 1995). Below is a brief overview of feed forward network for each training sample *X* and for each hidden or output layer node *j*:

- Calculate input I_j to that node as

$$I_j = \sum_{i=1}^{d} w_{ji} x_i + w_{j0} .$$

- Calculate output O_j from that node as

$$O_j = \frac{1}{1 + e^{-I_j}} .$$

Back-propagation is widely used algorithm for the purpose of training the neural networks. Back-Propagation algorithm can be considered as a two-step process. In the first step, the derivatives of the error function with respect to the weights are evaluated. In second step, these derivatives are then used to compute the adjustments to be made to the weights by using gradient descent or any other optimization schemes. An overview of back-propagation can be given as below:

- For each node *j* in output layer, calculate the error as

$$Err_j = O_j(1 - O_j)(T - O_j)$$

- For each node *j* in hidden layer, calculate the error as

$$Err_j = O_j(1 - O_j)\left(\sum_k Err_k w_{jk}\right)$$

- For each weight w_{ij}, calculate weight increment as

$$\Delta w_{ij} = l \cdot Err_j \cdot O_i$$

- Now, update the previous weight as

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

- For each bias θ calculate bias increment as

$$\triangle \theta_j = l \cdot Err_j$$

- Finally update bias with

$$\theta_j = \theta_j + \triangle \theta_j .$$

Neural networks are widely used in classification but there are still many unsolved issues in applying neural networks such as scalability, misclassification, convergence, higher mean square errors, and so forth. Many researchers have tried to overcome these issues and proposed variety of neural networks with better performances. Jiang, Harvey, and Wah (2003) proposed a new approach of constructing and training neural networks to overcome the problems including local minima and the slow convergence of the learning process. In their approach, feed-forward network was constructed based on the data clusters generated based on locally trained clustering (LTC) and then further trained using standard algorithms operating on the global training set which converges rapidly due to its inherited knowledge with good generalization ability from the global training. Ji and Ma (1997) proposed a learning method based on combination of weak classifiers, which were found by a randomized algorithm, to achieve good generalization and fast training time on both the test problems and the real applications. They showed that if the weakness factor was chosen according to the critical value given by their theory, the combinations of weak classifiers could achieve a good generalization performance with polynomial space- and time-complexity. Yu, Chen, and Cheng (1995) proposed dynamic learning using derivative information instead of fixed learning rate to optimize back propagation. The information gathered from the forward and backward propagation was used for dynamic learning; with this technique they achieved higher convergence rate and significant reduction in learning process. The probability of misclassification of any random sample can be termed as the generalization error of a classifier. Many researchers have used ensemble methods to reduce misclassification or generalization errors (Hansen & Salamon, 1990; Hashem & Schmeiser, 1995). Further information on classification errors, learning and generalization, and some of the recent developments in neural networks can be found in Kulkarni, Lugosi, and Venkatesh (1998), Solazzia and Uncinib (2004), and Zhang (2000). Recent studies in data mining techniques (Wu et al., 2008) do not consider ANN as a top 10 data mining technique.

SUPPORT VECTOR MACHINES

Support Vector Machine (SVM) is a state of the art machine learning algorithm (Cortes & Vapnik, 1995; Vapnik, 1995). The main idea of SVM is to separate the input space in two half spaces using the hyperplane $x^T w - y = 0$ which maximizes the margin between the two classes $\{A,B\}$. In other words, given n points $x_i \in R^d$, i=1,2,…,n, and the corresponding labels

$$y^i = \begin{cases} 1 \text{ if } x^i \in A \\ -1 \text{ if } x^i \in B \end{cases},$$

we need to find two parallel hyperplanes $x^T w - y = \pm 1$ which separates the points between two classes. The margin between two classes is represented by the distance between the two hyperplanes,

$$\frac{2}{\| w \|}.$$

Thus the optimization problem is to minimize the norm of w, with constraints to correctly classified points of both classes:

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 \\ &s.t.\ y^i (x^{i^T} w - y) \geq 1 \\ &\quad i = 1,2,\ldots,n \end{aligned}.$$

Here we note that minimizing $\| w \|$ is the same as minizing $\| w \|^2$. This method works when there is a perfect linear separation between the two classes. Therefore it is called the hard margin classifier. When the two classes are not linearly separable, the soft margin classifier is used. This classifier finds a hyperplane that allows a few points to violate the separation. The number of points that are not correctly classified should be minimized in this case.

Then the optimization problem is reduced to the following:

$$\min \frac{1}{2} \| w \|^2 + \frac{C}{2} \sum_{i=1}^{n} \xi^{i^2}$$
$$s.t. y^i (x^{i^T} w - y) \geq 1 - \xi^i$$
$$i = 1,2,...,n \quad .$$

These problems belong to the quadratic convex programming problems and can be solved using standard quadratic programming techniques (Bertsekas, 1995). For nonlinear classification, the SVM is used with kernel functions (Mangasarian & Wild, 2006) and the basic solution technique is through linear programming.

GENERALIZED EIGENVALUE CLASSIFICATION

Generalized eigenvalue classification methods proposed by Mangasarian and Wild (2004) (Guarracino, Cifarelli, Seref, & Pardalos, 2007) have a classification accuracy comparable with those obtained by SVM at a lower execution time. This is also a support vector machine classification where each plane of two non-parallel planes is generated such that it is closest to one of the two data sets and as far as possible from the other data set. This problem leads to two simple generalized eigenvalue problems. The main idea is to classify two sets of points, *A* and *B*, using two hyperplanes, each closest to one set of points, and furthest from the other. Let $x^T w - y = 0$ be a hyperplane in R^d . To satisfy the previous condition for the points in *A*, the hyperplanes can be obtained by solving the following optimization problem:

$$\min_{w,y \neq 0} \frac{\| Aw - ey \|^2}{\| Bw - ey \|^2} \tag{8}$$

The hyperplane for the *B* can be obtained by minimizing the inverse of the objective function in (8). If we use the notations $G = [A, -e]^T [A, -e]$, $H = [B, -e]^T [B, -e]$, $z = [w^T, y]^T$, then equation (8) becomes:

$$\min_{z \in R^d} \frac{z^T G z}{z^T H z} \tag{9}$$

The expression in (9) is the Raleigh quotient of the generalized eigenvalue problem $Gx = \lambda Hx$. The stationary points are obtained only at the eigenvectors of (9), where the value of the objective function is given by the eigenvalues. When *H* is positive definite, the Raleigh quotient is bounded and it ranges over the interval determined by minimum and maximum eigenvalues (Parlett, 1998). *H* is positive definite under the assumption that the columns of [*B, - e*] are linearly independent. The inverse of the objective function in (4) has the same eigenvectors and reciprocal eigenvalues. Let $z_{min} = [w_1, y_1]$ and $z_{\max} = [w_2, y_2]$ be the eigenvectors related to the eigenvalues of smallest and largest modulo, respectively. Then $x^T \ w_1 - y_1 = 0$ is the closest hyperplane to the set of points in *A* and the furthest from those in *B* and $x^T \ w_2 - y_2 = 0$ is the closest hyperplane to the set of points in *B*, and the furthest from those in *A*. In order to regularize the problem, we can solve:

$$\min_{w,y \neq 0} \frac{\| Aw - ey \|^2 + \delta \| \widetilde{B} \|^2}{\| Bw - ey \|^2 + \delta \| \widetilde{A} \|^2},$$

where the $\widetilde{A}$ and $\widetilde{B}$ are the diagonals of matrices $[Aw - ey]$ and $[Bw - ey]$, respectively.

By choosing the eigenvectors related to the new minimum and maximum eigenvalue, we obtain solutions that are close to the ones of the original problem. A point is classified according to the closest hyperplane or the class.

More recently, Cifarelli, Guarracino, Seref, Cuciniello, and Pardalos (2007) introduced IReGEC, an incremental technique capable of reducing the training set in the learning phase of the supervised classification, with the advantage of lower overfiting of data and improved classification accuracy. This method together

with some other classification methods has been applied to different biomedical data sets (Felici, Bertolazzi, Guarracino, Chinchuluun, & Pardalos, 2009; Guarracino et al., 2007). Incremental subset selection permits to construct a small set of points that retains the information of the entire training set and provides comparable accuracy results. A kernel built from a smaller subset is computationally more efficient in predicting new elements, compared to the one that uses the entire training set. Furthermore, a smaller set of points reduces the probability of overfitting the problem. Finally, as new points are available, the cost to retrain the algorithm decreases if the influence of those new points on classification is only evaluated with respect to that subset, rather than the whole training set.

APPLICATIONS

Historically one of the first papers dealing with data discrimination into different classes was directly applied to an agricultural problem. This is Fisher's Linear Discriminant Analysis (LDA) algorithm published in the annals of Eugenics (later renamed Annals of Human Genetics) and had been applied in order to discriminate 150 flowers into three classes of flowers based on four quantitative features namely: sepal width, sepal length, pedal width and pedal length (Fisher, 1936). Today this famous dataset (best known in data mining community as IRIS dataset) can easily be found in any open dataset repositories (e.g., UCI, http://archive.ics.uci.edu/ml/) and serves as an example in many undergraduate data mining courses around the world.

Although the first use data mining techniques was in agriculture, mathematical tools were not extensively used in this field for many years. The change came especially due to recent technological advances that made it possible to store and process large amount of data in home computers. Accumulation of data generated the challenge and the need for processing and analysis generating several well defined mathematical problems. Today many data miners identify and attack problems coming from agricultural science.

Clustering for example has been used (and more especially the famous *k-means* method) to address problems that arise during the fermentation process of wine. In Urtubia, Perez-Correa, Meurens, and Agosin (2004), the authors used the k-means to predict how good the fermentation process will be. This was achieved by recording a number of features related to sugars, alchohols organic acids and nitrogen sources. Initially, PCA was used to reduce the dimensionality of the problem and then the k-means clustering method was applied to determine the different clusters.

The k-means algorithm also has been used for image segmentation in the area of machine vision. As part of it, Leemans and Destain used k-means method for grading apples and identifying visually defected products before shipping them to the end customers (2004). They describe a conveyer based system able to analyze four apples in 1 second. The same authors extended their machine vision-based data mining using other methods such as the Linear Discriminant Analysis (Leemans, Magein, & Destain, 2002) and Bayesian (Leemans, Magein, & Destain, 1999) to solve the same problem.

On the other side Artificial Neural Networks (ANN) that are popular tools for especially among computer scientists for supervised classification problems. ANN's have demonstrated good results in practice, especially if the testing dataset is wisely chosen and the structure of the network is such that it is robust against overfitting. In Aerts, Jans, Halloy, Gustin, and Berckmans (2004), the authors utilize ANNs to detect abnormal coughing sounds in a herd. Abnormal coughing sound usually is associated with some disease so it is very crucial for the farmer to distinguish if there is any potentially diseased animal in the herd. Thus, data mining can be used to prevent and control the spread of dangerous contagious diseases.

Another application of ANN's in agriculture was presented by Shahin, Tollner, and McClendon (2001) where they are used as intelligent classifiers on Magnetic Resonance

Imaging (MRI) scans of apples in order to detect internal defects in apples (Shahin et al., 2001). From this perspective, neural network classification is used as a quality control tool to decrease the number of defected products. By utilizing X-ray images Schatzki, Haff, Young, Can, Le, and Toyofuku were able to detect about 90% of the defected apples (1997).

Liu, Lyon, Windham, Lyon, and Savage (2004) use PCA to analize chicken breast quality. In order to analyze the meat quality and the deboning time, a set of 36 chicken carcasses has been considered and randomly divided into four subgroups, each one containing 9 carcasses. These subgroups are designed for different deboning times. Chickens in the different groups have been deboned after 2, 4, 6 and 24 hours after the death. After deboning, the breasts have been cut in two parts, and each part has been subject to a different set of analysis. After using the PCA to decrease the dimensionality of the problem, results show that the first seven principal components are able to represent about the 70% percent of the total variations on the data. Moreover, the first four principal components represent about 50% of the total variations. In particular, the first principal component takes 23.3% of the variations, the second one 13.6%, the third one 8.8% and finally the fourth one 6.9%. An analysis on the data showed that the first component was mainly defined by the shear force and by the attributes decided by the group of panelists supervising this process. Therefore, these attributes are the most important variables for the evaluation of the chicken breast quality (Mucherino et al., 2009).

In Jagtap, Jones, LaRow, Ajayan, and O'Brien (2006) a *k*-NN algorithm is used for the recalibration of the precipitation outputs from the FSU-GSM (Florida State University *Global* Spectral Model) and FSU-RSM (Florida State University *Regional* Spectral Model) climate models. These climate models may not produce sufficiently accurate daily weather variable outputs to use in crop models. The objective is to find *k* neighboring years which have the forecasts closest to those of a target year. It is therefore assumed that the climate during a target year is a replication of the weather recorded in the past. The *k*-NN method resulted able to improve the accuracy of the monthly precipitation forecasts across all sites used in the study.

FUTURE DIRECTIONS

Data mining and knowledge discovery techniques are relatively new to agricultural and environmental fields. Their use is associated and conditioned with the use of research operations sets of tools. There are a number of research papers that show that agricultural and environmental sciences can really benefit from the use of mathematical tools and modern technology. It is important to note that the a number of published papers are purely research and have not yet been applied to be part of the set of tools farmers or practitioners use everyday. As an example, the study that use a Artificial Neural Network (Aerts et al., 2004) describes the process of how a pig goes through the system designed to record pig coughs to discover whether the animal has health problems but it does not address the issue of scaling the proposed system to be applied to the entire herd. The proposed system is complex and can examine only one animal at a time. It will be difficult to see the proposed system be applied in a herd of hundreds of pigs where each animal must be examined individually. The difficulties would be operational and finacial.

Often occurs that producers and practitioners in agriculture and environment have more trust on traditional decision making methods and they might see suspiciously automated decision making tools. This phenomenon has in general happened in all areas where data mining and knowledge discovery techniques have been introduced.

The answer that data miners provide is that classification, clustering and generally any other mathematical tool, doesn't aim to replace the human expert. On the contrary, they serve as expert assisting tools that help humans make sounder and better decisions. Data mining and

Knowledge discovery are totally human activities. Algorithms would provide the results that need to be interpreted by data miners in order to reach a final conclusion about the problem under study.

Many methods such as classical k-means clustering, ANNs, support vector machines, k-NNR classification have been applied in a variety of problems. But still there are many state of the art methods whose potential applications in agriculture have not been explored in full yet (e.g., biclustering). We envision more and more data mining application related papers with different problems and different algorithms being published in journals of interest to agricultural/environmental related scientists and practitioners.

We think that it is the time to introduce data mining and knowledge discovery techniques in the curriculum of agricultural and environmental departments, thus students can become familiar with these promissing techniques. The recently published book titled "Data Mining in Agriculture" (Mucherino et al., 2009) specifically tailored for students of agricultural and environmental fields will be a useful source of knowledge as it provides an exhaustive inventory of data mining and knowledge discovery techniques applied in agricultural and environmental sciences.

A number of academic institutions are organizing "Summer Schools" where in a week or two they provide a good introduction in data mining and knowledge discovery techniques. More of similar efforts will certainly help in making data mining and knowledge discovery techniques familiar to students, researchers and practitioners in the field of agriculture and environment.

REFERENCES

Aerts, J.-M., Jans, P., Halloy, D., Gustin, P., & Berckmans, D. (2004). Labeling of cough data from pigs for on-line disease monitoring by sound analysis. *American Society of Agricultural and Biological Engineers*, *48*(1), 351–354.

Arulselvan, A., Baourakis, G., Boginski, V., Korchina, E., & Pardalos, P. M. (2008). Analysis of food industry market using network approaches. *British Food Journal*, *110*(9), 916–928. doi:10.1108/00070700810900611

Bermejo, S., & Cabestany, J. (2000). Adaptive soft k-nearest-neighbour classifiers. *Pattern Recognition*, *33*, 1999–2005. doi:10.1016/S0031-3203(99)00186-7

Bertsekas, D. P. (1995). *Nonlinear programming*. Belmont, MA: Athena Scientific.

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.

Bishop, M. C. (2006). *Pattern recognition and machine learning*. New York: Information Sciences and Statistics, Springer.

Bradley, S., & Fayyad, M. (1998, July 24-27). Refining initial points for k-means clustering. In J. Shavlik (Ed.), *Proceedings of the 15th International Conference on Machine Learning (ICML98),* Madison, WI (pp. 91-99). San Francisco: Morgan Kaufmann.

Chinchuluun, R., Won Suk, L., Bhorania, J., & Pardalos, P. M. (2009). Clustering and classification algorithms in food and agricultural applications: A survey. In P. Papajorgji and P. M. Pardalos (Eds.), *Advances in modeling agricultural systems* (pp. 1-22). New York: Springer.

Cifarelli, C., Guarracino, M. R., Seref, O., Cuciniello, S., & Pardalos, P. M. (2007). Incremental classification with generalized eigenvalues. *Journal of Classification*, *24*(2), 205–219. doi:10.1007/s00357-007-0012-z

Cortes, C., & Vapnik, V. (1995). Support vector machines. *Machine Learning*, *20*, 273–279.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: John Wiley & Sons.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: John Wiley & Sons.

Dunn, J. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. doi:10.1080/01969727308546046

Felici, G., Bertolazzi, P., Guarracino, M. R., Chinchuluun, A., & Pardalos, P. M. (2008. November). Logic formulas based knowledge discovery and its application to the classification of biological data. In R. P. Mondaini (Ed.), 8th International Symposium on Mathematical and Computational Biology (*BIOMAT 2008),* Sao Paulo, Brazil (pp. 223-234). World Scientific.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Guarracino, M. R., Chinchuluun, A., & Pardalos, P. M. (in press). Decision rules for efficient classification of biological data. *Optimization Letters*.

Guarracino, M. R., Cifarelli, C., Seref, O., & Pardalos, P. M. (2007). A classification algorithm based on generalized eigenvalue problems. *Optimization Methods and Software*, *22*(1), 73–81. doi:10.1080/10556780600883874

Hammah, R. E., & Curran, J. H. (2000). Validity measures for the fuzzy cluster analysis of orientations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1467–1472. doi:10.1109/34.895981

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. doi:10.1109/34.58871

Hashem, S., & Schmeiser, B. (1995). Improving model accuracy using optimal linear combinations of trained neural networks. *IEEE Transactions on Neural Networks*, *6*(3), 792–794. doi:10.1109/72.377990

Hathaway, R., & Bezdek, J. (2001). Fuzzy *c*-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics - Part B, 31*(5), 735-744.

Hathaway, R., Bezdek, J., & Hu, Y. (2000). Generalized fuzzy c-means clustering strategies using L norm distances. *IEEE transactions on Fuzzy Systems*, *8*(5), 576–582. doi:10.1109/91.873580

Hung, M., & Yang, D. (2001, November 29-December 1). An efficient fuzzy c-means clustering algorithm. In *Proceedings of the IEEE International Conference on Data Mining,* San Jose, CA (pp. 225-232).

Hwang, W. J., & Wen, K. W. (1998). Fast k classification algorithm based on partial distance search. *Electronics Letters*, *34*(21), 2062–2063. doi:10.1049/el:19981427

Jagtap, S. S., Jones, J. W., LaRow, T., Ajayan, A., & O'Brien, J. J. (2006). (Manuscript submitted for publication). Statistical recalibration of precipitation outputs from coupled climate models. *Journal of Applied Meteorology*.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. doi:10.1145/331499.331504

Ji, C., & Ma, S. (1997). Combinations of weak classifiers. *IEEE Transactions on Neural Networks*, *8*(1), 32–42. doi:10.1109/72.554189

Jiang, X., Harvey, A., & Wah, K. S. (2003). Constructing and training feed-forward neural networks for pattern classification. *Patt. Recog.*, *36*, 853–867. doi:10.1016/S0031-3203(02)00087-0

Karayiannis, N. B. (1997). A methodology for constructing fuzzy algorithms for learning vector quantization. *IEEE Transactions on Neural Networks*, *8*(3), 505–518. doi:10.1109/72.572091

Krishna, K., & Murty, M. (1999). Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part B, 29*(3), 433-439.

Kulkarni, S. R., Lugosi, G., & Venkatesh, S. S. (1998). Learning pattern classification - a survey. *IEEE Transactions on Information Theory*, *44*(6), 2178–2206. doi:10.1109/18.720536

Kuncheva, L. I. (1997). Fitness functions in editing *k*-NN reference set by genetic algorithms. *Patt. Recog.*, *30*(6), 1041–1049. doi:10.1016/S0031-3203(96)00134-3

Leemans, V., & Destain, M. F. (2004). A real time grading method of apples based on features extracted from defects. *Journal of Food Engineering*, *61*, 83–89. doi:10.1016/S0260-8774(03)00189-4

Leemans, V., Magein, H., & Destain, M.-F. (1999). Defect segmentation on 'Jonagold' apples using colour vision and Bayesian method. *Computers and Electronics in Agriculture*, *23*, 43–53. doi:10.1016/S0168-1699(99)00006-X

Leemans, V., Magein, H., & Destain, M.-F. (2002). On-line apple grading according to European standards using machine vision. *Biosystems Engineering*, *83*(4), 397–404. doi:10.1006/bioe.2002.0131

Liu, Y., Lyon, B. G., Windham, W. R., Lyon, C. E., & Savage, E. M. (2004). Principal component analysis of physical, color, and sensory characteristics of chicken breasts deboned at two, four, six, and twenty-four hours postmortem. *Poultry Science*, *83*, 101–108.

MacQueen, J. B. (1967). Some methods for classification and analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability: Vol. 1* (pp. 281-297). Berkeley, CA: University of California Press.

Mangasarian, O. L., & Wild, E. W. (2004). *Multisurface proximal support vector classification via generalized eigenvalues* (Techn. Rep. 04-03). Madison, WI: Data Mining Institute, Computer Science Department, University of Wisconsin.

Mangasarian, O. L., & Wild, E. W. (2006). *Nonlinear knowledge-based classification* (Tech. Rep. 06-04). Madison, WI: Data Mining Institute, Computer Science Department, University of Wisconsin.

Marchant, J. A., & Onyango, C. M. (2003). Comparison of a Bayesian classifier with a multilayer feed-forward neural network using the example of plant/weed/soil discrimination. *Computers and Electronics in Agriculture*, *39*, 3–22. doi:10.1016/S0168-1699(02)00223-5

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Mucherino, A., Papajorgji, P., & Pardalos, P. M. (2009). Survey of Data Mining Techniques Applied to Agriculture. *Operational Research - . International Journal (Toronto, Ont.)*, *9*, 121–140.

Mucherino, A., Papajorgji, P., & Pardalos, P. M. (2009). *Data mining in agriculture*. New York: Springer.

Pan, J. S., Qiao, Y. L., & Sun, S. H. (2004). A fast k nearest neighbors classification algorithm. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences . E (Norwalk, Conn.)*, *87-A*(4), 961–963.

Parlett, B. N. (1998). The symmetric eigenvalue problem. *SIAM, 20,* 357.

Pernkopf, F. (2005). Bayesian network classifiers versus selective *k*-NN classifier. *Pattern Recognition*, *38*(1), 1–10. doi:10.1016/j.patcog.2004.05.012

Pyle, D. (2003). *Business modeling and data mining*. San Francisco, CA: Morgan Kaufmann Publishers.

Schatzki, T. F., Haff, R. P., Young, R., Can, I., Le, L.-C., & Toyofuku, N. (1997). Defect detection in apples by means of x-ray imaging. *Transactions of the American Society of Agricultural Engineers*, *40*(5), 1407–1415.

Shahin, M. A., Tollner, E. W., & McClendon, R. W. (2001). Artificial intelligence classifiers for sorting apples based on watercore. *Journal of Agricultural Engineering Research*, *79*(3), 265–274. doi:10.1006/jaer.2001.0705

Solazzia, M., & Uncinib, A. (2004). Regularizing neural networks using flexible multivariate activation function. *Neural Networks*, *17*(2), 247–260. doi:10.1016/S0893-6080(03)00189-8

Urtubia, A., Perez-Correa, J. R., Meurens, M., & Agosin, E. (2004). Monitoring large scale wine fermentations with infrared spectroscopy. *Talanta*, *64*(3), 778–784. doi:10.1016/j.talanta.2004.04.005

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., & Motoda, H. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *6*(14), 1–37. doi:10.1007/s10115-007-0114-2

Yager, R. R. (2006). An extension of the naive Bayesian classifier. *Information Science*, *176*(5), 577–588. doi:10.1016/j.ins.2004.12.006

Yager, R. R., & Filev, D. P. (1994). Approximate clustering via the mountain method. *IEEE Transactions on Systems, Man, and Cybernetics*, *24*(8), 1279–1284. doi:10.1109/21.299710

Yu, X., Chen, G., & Cheng, S. (1995). Dynamic learning rate optimization of the backpropagation algorithm. *IEEE Transactions on Neural Networks*, *6*(3), 669–677. doi:10.1109/72.377972

Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, *30*(4), 451–462. doi:10.1109/5326.897072

Zhang, Y., Xiong, Z., Mao, J., & Ou, L. (2006, June). The study of parallel k-means algorithm. In *Proceedings of the 6th World Congress on Intelligent Control and Automation: Vol.2,* Dalian, China (pp. 5868-5871). IEEE.

Altannar Chinchuluun received a BSc in mathematics in 2002 from the National University of Mongolia and a BSc in business administration in 2002 from the Mongolian University of Science and Technology. He received his MSc and a PhD in operations research at the University of Florida in December. Chinchuluun has received a number of prestigious awards such as Outstanding International Student Award, College of Engineering, University of Florida, 2007, Graduate Student Award for Excellence in Research, Department of Industrial and Systems Engineering, University of Florida, 2006, Outstanding International Student Award, College of Engineering, University of Florida, 2006, INFORMS 2006 Doctoral Colloquium participant, Department of Industrial and Systems Engineering, University of Florida, 2006, etc. Chinchuluun has published a number of papers in the most well-known international journals and edited several books. He is associate editor, Journal of Global Optimization*, Springer, guest editor, special issues on* Optimization and Optimal Control, Optimization*, Taylor & Francis, 2009. Currently he is a post doctoral associate at the Imperial College London, UK.*

Petros Xanthopoulos received the Dipl. Eng. degree in electronics and computer engineering from the Electronics and Computer Engineering Department, Technical University of Crete, Chania, Greece, in 2005 and a MSc from industrial and systems engineering, University of Florida. He is currently working toward the PhD degree in the Industrial and Systems Engineering Department, University of Florida, Gainesville. He has published his work in journals of IEEE, Elsevier and Wiley and he has edited a special issue for the Journal of Combinatorial Optimization *entitled* Data Mining in Biomedicine*. Currently he is editing a book on computational neuroscience that will be published from Springer. He has also co-organized (with Dr. Panos M. Pardalos) many special sessions in international meeting and two conferences related to applications of data mining in biomedicine. Petro Xanthopoulos has a patent with well-known researchers titled Time Frequency Transformation Analysis for Detection and Quantification of Epileptiform Activity Load in Generalized Epilepsies UF-731P. He is currently a research assistant at the Center for Applied Optimization, University of Florida.*

Vera Tomaino received the MS degree in industrial engineering from the University of Calabria, Italy, in 2007. She is currently pursuing her PhD in the Department of Biomedical Engineering at the University Magna Græcia, Catanzaro. Since November 2008, she is working as a visiting scholar at the Center for Applied Optimization, University of Florida. Her current research interests include data mining in biomedical applications and optimization. She is currently working on data mining applications in the field of biomedicine particularly applied to cancer research data. Tomaino has published her work in well-known journals and in international conferences. She closely works with the medical industry and research labs dealing with cancer research data applies data mining techniques and optimization.

Panos M. Pardalos is a distinguished professor of Industrial and Systems Engineering Department at the University of Florida and director of the Center for Applied Optimization (CAO). Professor Pardalos is editor in chief of five internationally well-known journals and member of the editorial board of more than 15 international journals in the field of applied mathematics. He is member of Academy of Sciences of Ukraine, Spain, Russia and Lithuania and Mongolia. Professor Pardalos is winner of The William Pierskalla best paper award for research excellence in health care management science. He has a long list of publications and books.