Combined supervised and unsupervised learning for system diagnosis using Dempster-Shafer theory *

Thierry Denœux and Gérard Govaert

Université de Technologie de Compiègne
U.R.A. CNRS 817 Heudiasyc
BP 529 F-60205 Compiègne cedex - France
email: Thierry.Denoeux,Gerard.Govaert@hds.utc.fr

ABSTRACT

A pattern recognition approach to system diagnosis is presented. This approach is based on supervised and unsupervised procedures both relying on the Dempster-Shafer theory of evidence. These procedures are combined to form an adaptive diagnosis system allowing for pattern rejection and creation of new classes. An application to water quality monitoring is presented.

1. INTRODUCTION

In the pattern recognition approach to system diagnosis, a feature vector is constructed using available information and classified into one of predefined categories corresponding to normal and abnormal system states [4]. However, conventional statistical techniques usually assume complete knowledge of all possible classes, a condition which is rarely satisfied in diagnosis applications. Consequently, a lot of research has been devoted to the development of adaptive pattern recognition systems allowing for pattern rejection and creation of new classes [4]. Such systems generally combine mechanisms for clustering similar patterns and decision rules for class assignment or rejection. Among the desired features of these systems are fast adaptation and learning, and efficient representation and management of uncertainty. In this paper, the Dempster-Shafer (D-S) theory of evidence is shown to provide a convenient framework for the development of a pattern recognition based diagnosis system.

A scheme for applying the theory of belief function to supervised pattern recognition was initially proposed in [3]. In this paper, we present an *unsupervised* learning algorithm based on similar ideas. The fundamental concept is that of a *stable* partition. A partition of a given set of vectors is said to be stable for a classification rule $\mathcal{R}$ if each vector remains in its own class when classified according to $\mathcal{R}$. The proposed clustering algorithm is shown to generate partitions that are stable for the evidence theoretic k-NN rule.

Once supervised and unsupervised learning algorithms have been defined, they can be combined to form an adaptive pattern recognition system. This combination is relatively easy in our case since both methods rely on the same basic principles. The system starts with a given set of training patterns labeled according to classes corresponding to known system states. Based on this knowledge, a supervised classification rule assigns new feature vectors to classes or rejects them. Once a predefined number of vectors have been rejected, an unsupervised mechanism generates a new partition, possibly including new classes. This partition is defined in such a way that each pattern is correctly classified using all other patterns.

The paper is organized as follows. The D-S theory of evidence is first briefly recalled in Section 2 and its application to supervised pattern classification is presented in Section 3. Section 4 introduces the clustering procedure and presents a proof of its convergence. Section 5 contains a description of a complete diagnosis system combining supervised and unsupervised learning mechanisms, as well as a demonstration of its effectiveness using real data from an environmental monitoring application. Some final remarks conclude the paper.

2. OVERVIEW OF D-S THEORY

The D-S theory of evidence originated from the concepts of upper and lower probabilities induced by a multi-valued mapping, introduced by Dempster in the 60's [1]. Shafer [6] showed the advantages of using belief functions for representing someone's degrees of belief, and provided a first formalization of a theory of belief functions. A new interpretation and an axiomatic justification of this theory was recently proposed by Smets [7], who also clarified the links between representation of belief and decision making.

One of the main concepts underlying D-S theory is that of *basic belief assignment* (BBA). Let us consider

* In P. Borne et al. (Eds) *CESA'96 IMACS Multiconference, Computational Engineering in Systems Applications. Symposium on Control, Optimization and Supervision*, vol.1, 104–109, Lille, July 9-12, 1996.

a set Ω of mutually exclusive and exhaustive hypotheses about some problem domain called the "frame of discernment". A piece of evidence that influences our belief concerning the truth value of each hypothesis in Ω, can be represented by a basic belief assignment m, which is defined as a function from 2^Ω to $[0,1]$ verifying $m(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m(A) = 1$. The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of m. The quantity $m(A)$ for any $A \subseteq \Omega$ can be interpreted as the "mass of belief" that one is willing to commit to A (and to none of its subsets), given the available evidence. Situations of complete ignorance therefore correspond to the case where $m(\Omega) = 1$ (m is then said to be *vacuous*). The information contained in m can be represented in two alternative forms: a belief or credibility function Bel, and a plausibility function Pl defined respectively as $Bel(A) = \sum_{B \subseteq A} m(B)$ and $Pl(A) = 1 - Bel(\bar{A})$. A natural interpretation of $Bel(A)$ is that of a measure of one's degree of belief in A, while $1 - Pl(A)$ provides a measure of doubt in that hypothesis. The width of the interval $[Bel(A), Pl(A)]$ can be regarded as a measure of one's ignorance of the "probability" of A, although the theory need not assume the existence of any objective probability distribution, as shown in [7].

The next part of D-S theory indicates how the evidence coming from several independent sources should be combined. Two BBAs m_1 and m_2 induced by distinct pieces of evidence are said to be *combinable* if there exist at least two non-disjoint subsets B and C of Ω such that $m_1(B) > 0$ and $m_2(C) > 0$. The orthogonal sum of m_1 and m_2, noted $m = m_1 \oplus m_2$, is then defined as $m(\emptyset) = 0$ and

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)} \tag{1}$$

for $A \neq \emptyset$. The orthogonal sum is commutative and associative, yielding invariance of the result with respect to the order in which the items of evidence are taken in consideration and combined.

Conventional Bayesian decision analysis is based on expected utilities (or losses) and requires uncertainty to be represented by probability distributions. Although the concept of mathematical expectation can be generalized in the framework of D-S theory [1], probability distributions can be shown to be better suited to decision making [7]. For that reason, Smets introduced the concept of "pignistic" probability $BetP$ obtained by distributing the mass $m(A)$ equally to each $\omega \in A$, for all $A \subseteq \Omega$: $BetP(\omega) = \sum_{\omega \in A} \frac{m(A)}{|A|}$, where $|A|$ denotes the cardinality of A. In this model, belief functions are proposed as a formalism for representing and updating belief, whereas the associated pignistic probabilities are used to make decisions.

3. PATTERN CLASSIFICATION

Principle

A classification procedure based on Dempster-Shafer theory was proposed in [3]. This procedure is briefly recalled in the following. Detailed discussion concerning the performance of this method and its ability to deal with data whose class label is uncertain can be found in [3].

Assume that some pattern $\mathbf{x}$ has to be classified according to a set Ω of M classes $\{\omega_1, \ldots, \omega_M\}$, using a training set of N samples with known classification. Each training pattern $\mathbf{x}^i$ can be regarded as a piece of evidence that influences one's belief in the class membership of $\mathbf{x}$. This evidence may be represented by a BBA m^i over Ω with two focal elements: the class of $\mathbf{x}^i$, denoted as ω_q, and Ω. The fraction of the unit mass assigned to $\{\omega_q\}$ can reasonably be defined as a decreasing function of the distance d^i (according to some appropriate distance measure) between vectors $\mathbf{x}$ and $\mathbf{x}^i$:

$$\begin{aligned} m^i(\{\omega_q\}) &= \alpha \phi_q(d^i) && (2) \\ m^i(\Omega) &= 1 - \alpha \phi_q(d^i) && (3) \\ m^i(A) &= 0 \quad \forall A \in 2^\Omega \setminus \{\{\omega_q\}, \Omega\} && (4) \end{aligned}$$

where $0 < \alpha < 1$ is a constant, and ϕ_q is a monotonically decreasing function verifying $\phi_q(0) = 1$ and $\lim_{d \to \infty} \phi_q(d) = 0$. In [3], an exponential form was postulated for ϕ_q:

$$\phi_q(d^i) = \exp(-\gamma_q (d^i)^2) \tag{5}$$

γ_q being a positive constant associated to class ω_q. As a result of considering each training pattern in turn, we obtain N BBAs $m^1, \ldots, m^N$. These functions can then be combined using Dempster's rule (equation 1) to yield a single BBA m.

Note that those training patterns $\mathbf{x}^i$ which are very dissimilar from $\mathbf{x}$ have very little influence on m, since m^i is then close to the vacuous BBA. Consequently, it is sufficient to consider in the combination only the set $\Phi_k(\mathbf{x})$ of the k nearest neighbors of $\mathbf{x}$ in the training set. In [9], a learning algorithm was proposed for determining the parameters γ_q in equation 5 by optimizing an error criterion; this improvement was shown to yield most of the time a substantial improvement of classification accuracy. A neural network based on the same approach was presented in [2].

Decision analysis

Decisions regarding the classification of pattern $\mathbf{x}$ based on a BBA m can be made as follows. Let us denote as $\mathcal{A}$ a finite set of actions, including, e.g., assignment to each class and rejection. The loss incurred if one chooses action $\alpha \in \mathcal{A}$ whereas pattern $\mathbf{x}$ belongs to class $\omega \in \Omega$ is denoted by $\lambda(\alpha|\omega)$. Assuming the only focal elements of m to be singletons and Ω, the risk of choosing action α relative to the pignistic probability distribution is:

$$R(\alpha|\mathbf{x}) = \sum_{\omega \in \Omega} \lambda(\alpha|\omega) BetP(\{\omega\}) \quad (6)$$

$$= \sum_{\omega \in \Omega} \lambda(\alpha|\omega) \left(m(\{\omega\}) + \frac{m(\Omega)}{|\Omega|} \right) \quad (7)$$

The Bayes decision rule then prescribes the action for which the risk is the smallest. Let us now consider two special cases in greater detail.

Case 1: The frame Ω is composed of M classes $\{\omega_1, \ldots, \omega_M\}$ that are all represented in the training set. The actions are $\mathcal{A} = \{\alpha_1, \ldots, \alpha_M\}$, where α_i denotes assignment to class ω_i. The loss is 0 for correct classification and 1 for misclassification. We then have:

$$R(\alpha_i|\mathbf{x}) = \sum_{j \neq i} m(\{\omega_j\}) + \frac{M-1}{M} m(\Omega) \quad (8)$$

$$= 1 - m(\{\omega_i\}) - \frac{m(\Omega)}{M} \quad (9)$$

$$= 1 - BetP(\{\omega_i\}) \quad (10)$$

Hence, the Bayes rule results in this case in assignment of pattern $\mathbf{x}$ to the class with the largest pignistic probability (or, equivalently, with the largest mass of belief).

Case 2: The frame Ω is composed of M *known* classes $\{\omega_1, \ldots, \omega_M\}$ and one *unknown* class ω_u representing those states of nature that have not yet been encountered and for which, consequently, no example is available in the training set. In addition to assignment to all known classes, the set of possible actions now includes assignment to the unknown class, denoted by α_u, and rejection α_0. As before, the loss of wrongly assigning a pattern to one of the known classes is taken equal to 1; λ_0 and λ_1 denote the cost of rejection and misclassification in the unknown class, respectively; all other losses are assumed to be equal to 0. The different risks can then be computed as:

$$R(\alpha_i|\mathbf{x}) = 1 - BetP(\{\omega_i\}) \quad (11)$$

$$\text{for } i = 1, \ldots, M$$

$$R(\alpha_0|\mathbf{x}) = \lambda_0 \quad (12)$$

$$R(\alpha_u|\mathbf{x}) = \lambda_1 \left(1 - \frac{m(\Omega)}{M+1}\right) \quad (13)$$

$$= \lambda_1 (1 - BetP(\{\omega_u\}) \quad (14)$$

Examples of decision regions induced by the strategy of risk minimization in this case will be presented in section 5.

4. CLUSTERING PROCEDURE

Principle

In this section, we shall demonstrate how some of the ideas presented above may be extended to provide the basis for an unsupervised learning procedure. As before, we consider a set $\mathcal{X}$ of N patterns $\mathbf{x}^1, \ldots, \mathbf{x}^N$ that will be provisionally assumed to be partitioned in M classes $\omega_1, \ldots, \omega_M$. This partition of $\mathcal{X}$ is denoted by $\mathcal{P}$, and the class label of $\mathbf{x}^i \in \mathcal{X}$ is denoted as $L^i \in \{1, \ldots, M\}$. We denote by $\mathcal{R}$ the decision rule corresponding to case 1 in the above discussion. Assume that an arbitrary pattern $\mathbf{x}^i$ would be assigned to class $q = \mathcal{R}(\mathbf{x}^i)$ using the $N-1$ other training patterns. If $L^i = q$, then $\mathbf{x}^i$ will be said to be correctly classified by $\mathcal{R}$. If all patterns in $\mathcal{X}$ are correctly classified by $\mathcal{R}$, then partition $\mathcal{P}$ is *stable*. Henceforth, stability will be regarded as a desired feature for a partition of a data set.

We now consider the following procedure, denoted by Π. Starting from an arbitrary initial partition, each pattern $\mathbf{x}^i$ is selected in random order from $\mathcal{X}$ and classified by $\mathcal{R}$. If $\mathcal{R}(\mathbf{x}^i) = L^i$, then L^i remains unchanged, otherwise it is changed to $\mathcal{R}(\mathbf{x}^i)$. The same operation is repeated until a stable partition is obtained, i.e. no change in class label occurs. We shall now prove that this procedure terminates, i.e. a stable partition is always obtained in a finite number of iterations.

Proof of convergence

We first need to compute the mass assignment m^i obtained by considering all patterns $\mathbf{x}^j \in \mathcal{X}$ with $j \neq i$. Pattern $\mathbf{x}^j$ with $L^j = q$ yields a BBA $m^{i,j}$ computed using equations 2 to 3. For simplicity we assume $\phi_1 = \ldots = \phi_q = \phi$:

$$m^{i,j}(\{\omega_q\}) = \alpha\phi(d^{i,j}) \quad (15)$$

$$m^{i,j}(\Omega) = 1 - \alpha\phi(d^{i,j}) \quad (16)$$

and $m^{i,j}(A) = 0$ for all other $A \in 2^\Omega$, $d^{i,j}$ being the distance between $\mathbf{x}^i$ and $\mathbf{x}^j$. By applying equation 1, the combination of the $m^{i,j}$ for all $j \neq i$ can be shown [3] to lead to:

$$m^i(\Omega) = \frac{1}{K} \prod_{j \neq i} T_{i,j} \quad (17)$$

$$m^i(\{\omega_q\}) = \frac{1}{K} \left(\prod_{\{j \neq i, L^j \neq q\}} T_{i,j} - m^i(\Omega) \right) \quad (18)$$

for all $\omega_q \in \Omega$, K being a normalizing constant and $T_{i,j} = 1 - \alpha\phi(d^{i,j})$.

To demonstrate the convergence of procedure Π, we shall use an argument similar to that used by Hopfield in the analysis of dynamic recurrent neural networks [5], with which our approach has a formal analogy. We consider the following cost function:

$$E = \frac{1}{2}\sum_{q=1}^{M} \sum_{\{i \neq j | L^i = L^j\}} w_{i,j} \qquad (19)$$

where $w_{i,j} = -\ln T_{i,j}$. According to the terminology introduced by Shafer [6], $w_{i,j}$ is the weight of the evidence of pattern $\mathbf{x}^j$ relative to the class of pattern $\mathbf{x}^i$. Note that $w_{i,j} = w_{j,i}$. We shall show that each change in a class label strictly increases E. Since E can only take on a finite number of values, no change occurs after a finite number of iterations.

Assume that the class label of a pattern $\mathbf{x}^i$ has been changed from q to p. The corresponding change in E is:

$$\Delta E = \sum_{\{j \neq i | L^j = p\}} w_{i,j} - \sum_{\{j \neq i | L^j = q\}} w_{i,j} \qquad (20)$$

Now, given that L^i was changed from q to p, we necessarily had $m^i(\{\omega_p\}) > m^i(\{\omega_q\})$. It follows that:

$$\prod_{\{j \neq i, L^j \neq p\}} T_{i,j} > \prod_{\{j \neq i, L^j \neq q\}} T_{i,j} \qquad (21)$$

$$\Leftrightarrow \sum_{\{j \neq i, L^j = p\}} w_{i,j} > \sum_{\{j \neq i, L^j = q\}} w_{i,j} \qquad (22)$$

Hence $\Delta E > 0$, which completes the proof.

Implementation

In the above procedure, the computation of m^i requires $O(N)$ operations and must be repeated N times before all patterns are selected. Hence, the complexity is $O(N^2)$ for each complete sweep through the data set. This complexity can be significantly reduced if we classify each point with respect to its k nearest neighbors with $k < N$. Denoting by $\Phi_k(\mathbf{x}^i)$ the set of the k nearest neighbors of $\mathbf{x}^i$, m^i can then be defined as in equations 17 and 18, with now $T_{i,j} = 1 - \alpha\phi(d^{i,j})$ if $j \neq i$ and $\mathbf{x}^j \in \Phi_k(\mathbf{x}^i)$, and 1 otherwise. However, with this definition, matrix $(w_{i,j})$ is no longer symmetric and the above argument concerning the convergence of the algorithm is no longer valid (this condition was used to derive equation 20). To make matrix $(w_{i,j})$ symmetric, we can add to the k-NN of $\mathbf{x}^i$ those patterns of which $\mathbf{x}^i$ is one the k-NN:

$$\Phi'_k(\mathbf{x}^i) = \Phi_k(\mathbf{x}^i) \cup \{\mathbf{x} \in \mathcal{X} | \mathbf{x}^i \in \Phi_k(\mathbf{x})\} \qquad (23)$$

Another issue of great importance in practice is the determination of the number of classes. Since our algorithm in its present form can only decrease the initial number of classes, a strategy which has proved very efficient in practice consists in starting with a trivial partition for which there is one different class for each training pattern. With this initialization scheme, the number of classes has not to be defined in advance. With $\phi(d)$ defined as $\exp(-\gamma d^2)$, the only tuning parameters besides k are then α and γ, of which only the latter was found to have significant influence on the results.

5. APPLICATION TO DIAGNOSIS

An adaptive procedure

The above supervised and unsupervised procedure may be combined as follows to form an adaptive diagnosis system.

Assume that at time t our knowledge of the system under study consists in a set of N labeled measurement vectors representative of M known system states (corresponding, e.g., to normal conditions or different categories of defects). This training set is not necessarily supposed to be exhaustive, i.e., there may exist unknown states from which no training sample is available. As before, known states are denoted by $\omega_1, \ldots, \omega_M$ while unknown states are collectively designated by ω_u.

When classifying new measurement vectors, it is usually very important to recognize whether the current situation corresponds to one of the known classes or to an unknown class. When the uncertainty is too high, it may also be wise to postpone decision making until more information is available. We therefore place ourselves in the situation described as case 2 in Section 3. Following the terminology introduced by Dubuisson [4], actions α_u and α_0 will be called "distance" and "ambiguity", respectively.

Once the number of distance rejected observations potentially allows the characterization of a new state, we propose to start a clustering phase. This may be done using the procedure described in Section 4 operated in "semi-supervised mode", i.e., by clamping the class label of initial training patterns and adjusting the labels of rejected points until a stable partition of the data set is obtained. If M' new clusters have been identified, the set of classes is extended to include M' corresponding system states. Ambiguity rejected points are then assigned to new or old states according to the supervised decision rule without rejection (case 1 in Section 3).

Simulation results

To demonstrate the efficiency of the above procedure, we now consider real data from an environmental monitoring application [8]. The data consists of daily measurements of water quality parameters (pH, conductivity, NO3 and NH4 concentrations) performed in the river Seine during two years. The sampling point was located upstream a drinking water production plant. A monitoring system under development will process continuous measurements of the same parameters to produce in real-time a diagnosis of water quality. The data considered here has been used for prototyping of the system. For better visualization of the results, the data dimension was reduced to 2 by considering only pH and conductivity measurements.

The results obtained with our method at different time steps are illustrated in figures 1 (a) to (f). Ambiguity and distance rejected points are represented by + and o, respectively. The dotted lines delimit the regions of assignment to a known class; the boundaries of distance rejection regions are indicated as solid lines. The initial data consisted of the first 30 measurements. At time $t = 76$, 30 patterns have been distance rejected (a) and the clustering procedure is run, yielding one new class (b). At $t = 164$, 30 new patterns have again been rejected (c), but no class is created (d). A third class appears at $t = 186$ (e-f). The procedure was then iterated until $t = 731$ (end of second year) without creation of any new classes. Close examination of the results by domain experts led to the interpretation of the three clusters as corresponding to two nominal states and one particular form of pollution.

6. CONCLUDING REMARKS

A framework for supervised and unsupervised pattern classification based on the Dempster-Shafer theory of belief functions has been presented. This framework allows both rejection of atypical or ambiguous patterns and clustering of rejected points, which makes it well suited for diagnosis applications. While such an approach may be used routinely for monitoring a complex (technological or natural) system and its evolution, it must emphasized that some caution should be exercised when interpreting the data clusters exhibited by the system, which typically requires the assistance of a domain expert. In that sense, such a system cannot generally be expected to run for a long time without at least some kind of human supervision. Further work will aim at further investigating some theoretical aspects associated to uncertainty representation, and at applying this methodology to a full-scale application.

7. REFERENCES

[1] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, AMS-38:325–339, 1967.

[2] T. Denœux. An evidence-theoretic neural network classifier. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 712–717, Vancouver, October 1995.

[3] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[4] B. Dubuisson. *Diagnostic et Reconnaissance des Formes*. Hermès, Paris, 1990.

[5] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558, 1982.

[6] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, N.J., 1976.

[7] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–243, 1994.

[8] T. Trautmann, M. Clément, T. Denœux, and T. Wittig. Application of intelligent techniques to river quality monitoring. In *Proceedings of EUFIT'95*, volume 3, pages 1603–1609, Aachen, August 1995.

[9] L. M. Zouhal and T. Denœux. An adaptive k-NN rule based on Dempster-Shafer theory. In *Proc. of the 6th Int. Conf. on Computer Analysis of Images and Patterns (CAIP'95)*, pages 310–317, Prague, September 1995. Springer Verlag.

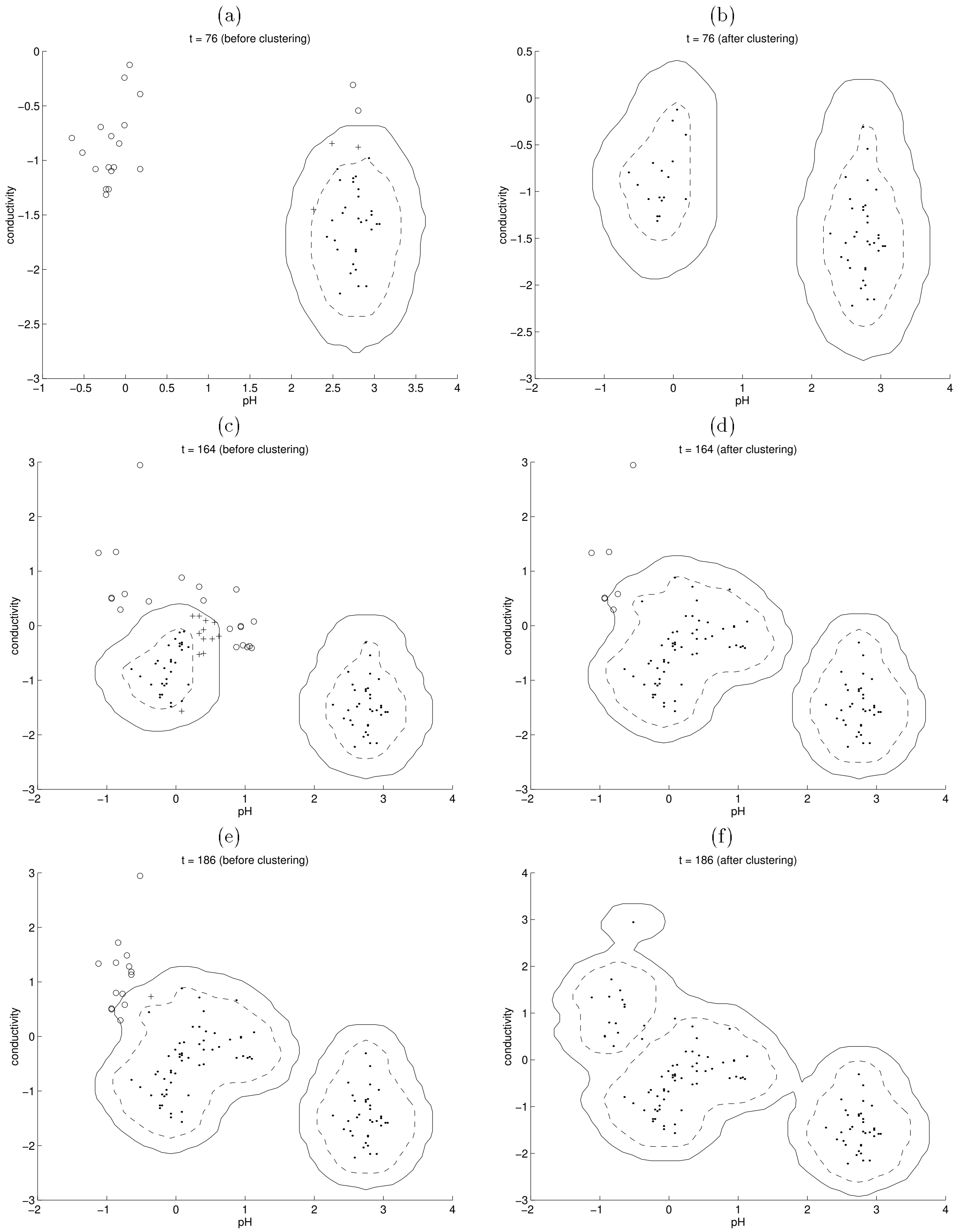

Figure 1: Simulation results with water quality data.