J Math Imaging Vis (2012) 42:212–224
DOI 10.1007/s10851-011-0315-x

Isotropic Image Analysis for Improving CBR Forecasting

Aitor Mata · M. Dolores Muñoz · Emilio Corchado · Juan M. Corchado

Published online: 1 September 2011

Abstract A novel hybrid forecasting Case-Based Reasoning (CBR) system is presented in this interdisciplinary study in which an isotropic buffer operator is applied for case-based creation. Commonly used as an image analysis technique by commercial Geographic Information Systems (GIS), the buffer operator in this particular system calculates the area of an oil slick for prediction and visualization tasks. The use of the buffer operator improves the quality of the data used by the system and in consequence the quality of the results obtained. The system generates predictions by using historical data on oil-slick formation following a spill.

Keywords Isotropic image analysis · Hybrid intelligent system · Case-Based Reasoning · Forecasting · Oil spill

1 Introduction

The emergency response to minimize the environmental impact when an oil spill occurs should be precise, fast and well-coordinated. The use of contingency response systems can facilitate planning and task allocation when organizing resources, especially when multiple teams and systems are deployed. This research presents a novel hybrid system for helping to manage such situations that involves a Case-Based Reasoning model to predict oil slick formation and the direction of oil slick drifts in a certain areas of the ocean. It also applies a buffer operator to calculate the size of the oil slicks in satellite images.

The term buffer operator, also known as the buffering or influence zone is used by all customized Geographical Information Systems (GIS). This operator is defined as the geometric space of the points that are at a shorter or similar distance to a given object (point, polyline or polygon) [12]. This definition is isotropic or directionally uniform, since the distance of the object to the edge of the buffer is constant in any direction on the plane. Among other fields, this operator is used in the simulated visualization of environmental processes, such as surveys of pesticide and chemical fertilizer contamination in shallow water tables in hydrographic basins; the influence of nitrates and silt levels on the growth of local flora, the environmental impact of installing new industries in close proximity to urban centers, the determination of areas of high seismic risk and so on.

There are two methods for the generation of influence areas: Voronoi triangulation and the Minkowski Sum [39]. In the latter method, a secondary polygon or generating polygon is defined as located on a point or moving on a polyline or polygon and generating a surface formed by the points over which the generating polygon moves. In the isotropic buffer, the generating polygon is a circle, which implies a constant distance between the border of the buffer and the object.

Case-Based Reasoning (CBR) [27] systems make use of past information in order to generate new solutions to new problems. The quality of the information stored within the case base will determine the quality of the solutions offered by these systems. Thus, the isotropic buffer operator is an important element in image analysis, and in this frame it pro-

A. Mata · M.D. Muñoz · E. Corchado (✉) · J.M. Corchado
University of Salamanca, Plaza de la Merced, s/n, Salamanca, Spain
e-mail: escorchado@usal.es

A. Mata
e-mail: aitor@usal.es

M.D. Muñoz
e-mail: mariado@usal.es

J.M. Corchado
e-mail: corchado@usal.es

vides the CBR system with accurate information that may be used in future situations.

The rest of the paper is organized as follows. The following section gives a brief explanation of CBR methodology. Section 3 develops the concept of isotropic image analysis, introducing the main terms of the directional statistics, after which Sect. 4 describes the use of the buffer operator in the existing GIS, and Sect. 5 provides an explanation of the CBR system presented in this study. The penultimate and final sections are respectively dedicated to the results obtained from applying the system to a real-life case study and the conclusions.

2 Case-Based Reasoning

Case-Based Reasoning is a technique that has its origin in knowledge-based systems. CBR systems learn from previous situations [1]. The main element of a CBR system is the *case base*; a structure that stores problems, elements (*cases*), and their solutions. So, a case base can be visualized as a database that stores a collection of problems with some sort of relationship to the solutions to every new problem, which gives the system the ability to generalize in order to solve new problems.

The learning capabilities of CBR systems rely on their own structures, which consist of four main phases [2]: *retrieval, reuse, revision* and *retention*. Figure 1 shows a graphical representation of those four phases. The *retrieval* phase consists of finding the cases in the case base that most closely resemble the proposed problem. Once a series of cases have been extracted from the case base, they must then be *reused* by the system. In this second phase, the selected cases are adapted to fit the current problem. After offering a solution to the problem, it is then *revised*, to check whether the proposed alternative is in fact a reliable solution to the problem. If the proposal is confirmed, it is *retained* by the system and could eventually serve as a solution to future problems.

Because it is a methodology [50], Case-Based Reasoning has been used to solve a great variety of problems. It is a cognitive structure that can be easily applied to solve problems such as those related to soft computing [37], since the procedures it uses are quite easy to assimilate in the soft-computing approaches. CBR has also helped to create applications for a variety of environments, such as health sciences [10, 38], where images can often play an important role [4, 24], or eLearning [3, 13]. As it has evolved, CBR has been used to solve new problems, applied as a methodology to create plans, and broken down into a distributed version [42]. Oceanographic problems [17] have also been addressed using these techniques in order to predict the value of highly inconsistent parameters.

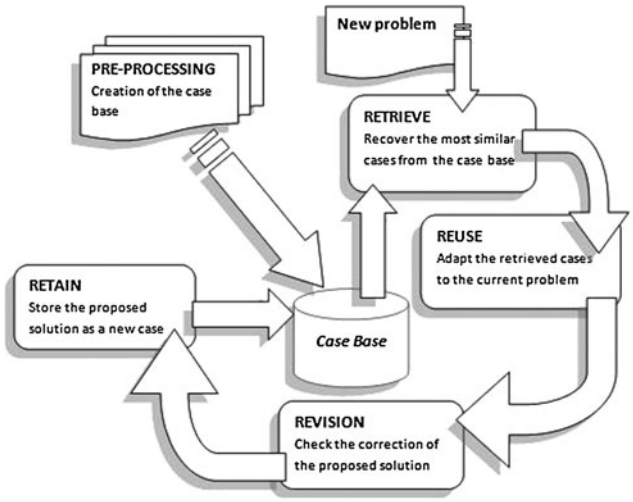

Fig. 1 Basic representation of the CBR cycle

The use of past knowledge to generate new solutions makes CBR systems very useful as decision support systems. Distributed and multi-agent [7] systems have used the CBR methodology to exploit its decision-support capabilities as an addition to their characteristics. On the other hand, as it is a methodology, CBR has been successfully applied to quite different knowledge fields and combined with a great variety of techniques. Most of the techniques used within CBR systems serve to classify, adapt, revise solutions, etc. Artificial neural networks such as ART-Kohonen neural networks and fuzzy logic have also been used to complement the capabilities of the CBR methodology [25]. Similarity measures, such as the $k-NN$ (k nearest neighbors) and also modern variations such as Significant Nearest Neighbor [49] where the value of k, which is the number of neighbors to consider, is calculated by taking into account the dissimilarity between the new case and the past cases stored in the case base, have been used to retrieve cases from the case base. Numeric situations, like those used in microarray problems, can be reused through neural networks, such as Growing Cells Structures (GCS) [14], where the aim is to cluster the retrieved information. Another way of using neural networks to adapt the retrieved information is to change the weight of the connection between the neurons depending on the retrieved cases [54]. Changing the weights allows the system to adapt the solution to the problem, as the retrieved cases will depend directly on the proposed problem. If the case-base structure is integrated into a neural network, then the revision phase consists of changing the organization of the case base, depending on the correction of the proposed result and other neural variables such as neuron age, activation value and last use [51]. Genetic algorithms (GA) are also used to revise the correction of the solutions [40]. After running those algorithms, the solutions can be accepted, and added to the case base.

Current trends in CBR are exploring the possibility of providing explanations from the actual CBR systems [47]. These techniques allow the CBR systems to give the users a better solution, adding additional information to the solution proposed by the system in the form of an explanation. With the explanations generated by the system, the solutions it proposes are justified and may be better understood.

Applying CBR to solve a problem generally implies complementing it with other artificial intelligence techniques. These models do not only provide simple ways of structuring the reutilization of the information, but they can also combine different techniques to improve on individual results. CBR has been used in combination with artificial intelligence techniques to boost the power of the core methodology [29]. Different kinds of neural networks, such as ART-Kohonen neural networks [20, 52, 53], have been used with CBR to create the internal structure of the case base automatically [7]. Even fuzzy logic [18] has been used to complement the capabilities of the CBR methodology. In this case, Growing Cell Structures will be used to structure the case base and to easily and accurately recover the most similar cases from the case base.

3 Directional Statistics

Directional statistics [19] is concerned with data that represent vectors on a plane or in 3D space. In the former case, the sample space is a circle and in the second, a sphere. In order to take the structure of these sampling spaces into account, special statistical methods are needed. Some examples of circular data are the directions of the predominant winds, the flight of migratory birds; indeed, any data that can be measured and converted into degrees or radius may be considered circular data.

Circular data may be represented as a point on a unit circle or as a unitary vector on the plane. Take a direction as the origin and choosing a direction, each circular datum may be specified as having an angle θ between the initial direction and the direction that corresponds to the datum (Fig. 2). The direction of the vector is characterized by the angle θ. The simplest way of representing circular data is to draw them as points on the unit circle and, when a direction repeats, place the new points outside the circle on the corresponding radius (Fig. 3).

We can define the circular variables [36] as those that represent directions on the plane, which is quantified by angles that vary from 0 to 2π. One of the most important differences in comparison with linear variables is that while these can take values of the whole straight line, circular variables take cyclic values and, consequently, the sum or difference of observations can surpass 360° or can even take a negative value; it being possible in these cases to find an equivalent

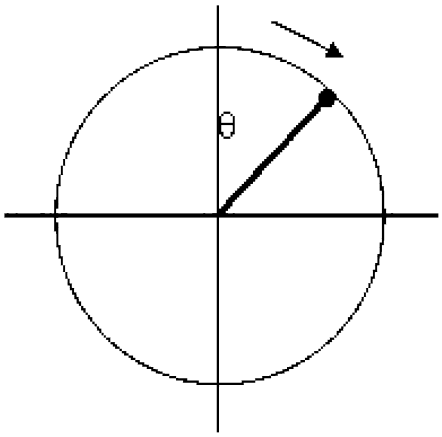

Fig. 2 Representation of a unitary vector on the plane on a circular graphic

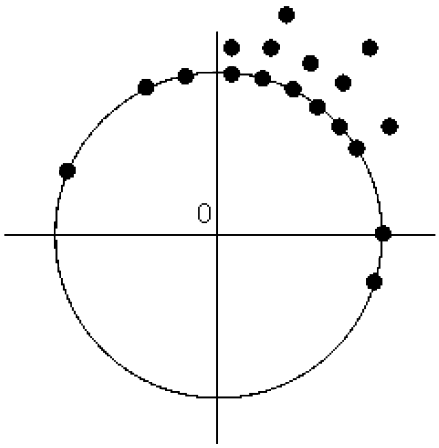

Fig. 3 Representation of circular data as points on the unit circle

value in the interval 0–360°. This characteristic means that circular variables may be treated in a different way to the linear ones, through statistical methods, correlation analysis and specific distributions for this type of variables.

Conceptually, a circular distribution can be considered a bivariated linear distribution where the total probability (or total mass) is dispersed around the circle unit. Therefore, as in the bivariated linear statistic, a mean vector $\bar{m}$ of module r and mean angle $\bar{\Phi}$ exists in circular statistics, at the tip of which is the mass center C of the distribution (Fig. 4).

3.1 Mean Vector

The mean statistical vector $\bar{m}$ is calculated by assigning to each of the points n of the unit radius circle in Fig. 5, a mass M, from which the centre of mass or gravity C is calculated.

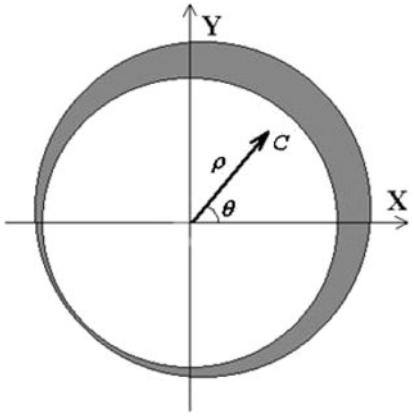

Fig. 4 Circular distribution with mean vector of module r and mean angle $\bar{\Phi}$. C is the mass centre of the distribution

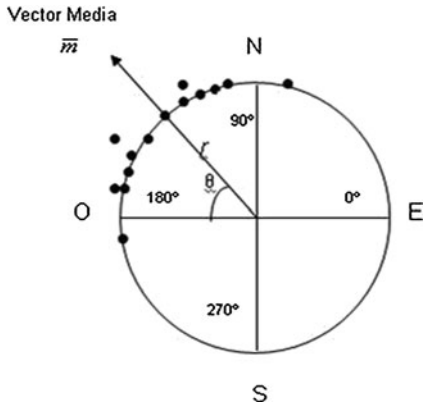

Fig. 5 Mean vector

The projections $\bar{x}$ and $\bar{y}$ of $\bar{m}$ are given by the expressions:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{j} n_i \operatorname{Cos}\Phi_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{j} n_i \operatorname{Sin}\Phi_i \tag{1}$$

$$n = \sum_{i=1}^{j} n_i$$

Let $Z(\Phi)$ be the random variable. If we take a monomodal sample of frequencies $n_1, n_2, \ldots, n_j$ in the directions $\phi_1, \phi_2, \ldots, \phi_j$, the mean vector $\bar{m}(r, \bar{\phi})$ is defined as

$$r = \sqrt{\bar{x}^2 + \bar{y}^2}$$

$$\bar{\Phi} = \begin{cases} \operatorname{Arctan}[\bar{y}/\bar{x}] & \text{if } \bar{x} > 0 \\ 180 + \operatorname{Arctan}[\bar{y}/\bar{x}] & \text{if } \bar{x} < 0 \end{cases} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the projections of $\bar{m}$ on the X and Y axes respectively. The direction of the vector is found on the straight line that joins C with the coordinates of the origin O. When the data are grouped in arcs j with a length of $\lambda = \frac{2\pi}{j}$, the values of $\bar{x}$ and $\bar{y}$ are as follows:

$$\bar{x} = \frac{1}{n}\left[n_1 \cos\Phi_1 + n_1 \cos\Phi_1 + \cdots + n_j \cos\Phi_j\right]$$

$$\bar{y} = \frac{1}{n}\left[n_1 \sin\Phi_1 + n_1 \sin\Phi_1 + \cdots + n_j \sin\Phi_j\right] \tag{3}$$

where, $n_1, n_2, \ldots, n_j$ are the frequencies of the mean points $\phi_1, \phi_2, \ldots, \phi_j$ of the arcs j. The value θ is estimated in the same way as in grouped samples. However, it needs a correction factor. Without correction, r tends to be a little smaller. Therefore, r has to be multiplied by a factor of $c > 1$. The corrected module is $r_c = r \bullet c$ where the correction factor c is:

$$c = \frac{\lambda/2}{\operatorname{Sin}(\lambda/2)} \qquad (\lambda \text{ in radians}) \tag{4}$$

The calculation described for $\bar{m}$ is valid for mono-modal samples. There exist many algorithms that allow to obtain the better fitting circles to the data graphically shown in Fig. 4. In [28], an analysis of currently used algorithms used to make that adjustement is done and a new one is proposed, based on left and right side partial derivatives. However, the experience shows that the phenomena linked to orographic discontinuities may be plurimodal. The process of v-modal samples (where v is the number of modes) differs from what is described, as these should be treated as if they were samples generated by v unimodal distributions, which is why we can talk about a mixture of distributions.

The v-modal samples should be considered as extracted from a distribution that is generated by the overlap of v monomodal distributions. When the distances between modes are arbitrary, no standard methods exist to breakdown a v-modal sample into v-mono-modal samples; in nature, plurimodal samples usually appear as bimodal and diametrically opposed. In this case, it is possible to reduce the bimodal sample to a monomodal sample by duplicating the angles. With the new angles, the mean vector $\bar{m}_2(r_2, \bar{\Phi}_2)$ is calculated by using (1)–(3). In order to obtain the symmetric modal angle $\bar{\Phi}_1$ from the original sample, we have to cancel out the effect of the duplication of the angles, as follows:

$$\bar{\Phi}_1 = \bar{\Phi}_2/2 \quad \text{or} \quad \bar{\Phi}_1 = \bar{\Phi}_2/2 + 180^\circ \tag{5}$$

3.2 Von Misses Distribution

Among the existing circular distributions [22, 46], one of the most widely used for the modeling of circular variables is the Von Misses distribution, in which the density function for v-modal and symmetric samples is

$$f(\Phi) = \frac{1}{2\Pi I_0(k)} \operatorname{Exp}[k \operatorname{Cos} v(\Phi - \Theta)] \tag{6}$$

where I_0 is the Bessel function of an imaginary pure argument of order 0, v is the number of modes, and k is the concentration parameter [44], that indicates the extent to which the distribution around the dominant direction Θ is concentrated.

For $k = 0$, $f(\Phi)$ degenerates in an uniform distribution. Mardia demonstrated [33] that the maximum likelihood estimation $\hat{\theta}$ and $\hat{\rho}$ for parameters θ and ρ of a Von Misses distribution are respectively $\bar{\Phi}$ and r. Likewise,

$$\frac{I_1(\hat{k})}{I_0(\hat{k})} = r \tag{7}$$

is fulfilled. Hence, the solution to (7) is the maximum likelihood $\hat{k}$.

3.3 Minkowski Sum

Given two images, A and B in R^2, the Minkowski sum is defined as [45].

$$A \oplus B := \bigcup_{b \in B} A + b \tag{8}$$

Where A is the generating polygon, and B the skeleton or primary element (point, polyline, or polygon). $A \oplus B$ is generated by moving A though each element $b \in B$, and then by adding the result of all the translations later on. The translation of the generating polygon A through the element $b \in B$ is defined as

$$A + b := \{a + b, a \in A\} \tag{9}$$

If we take a circle as the generating polygon A, and the group of points $B = \{(2, 3), (3, 4), (2, 5, (1, 5))\}$ as the primary element:

$$A \oplus B := [A + (2, 3)) \cup (A + (3, 4)) \cup (A + (2, 5)) \cup (A + (1, 5))] \tag{10}$$

Figure 6 shows the result, as well as $A \oplus L$ and $A \oplus P$, additions which have respectively taken polyline L and polygon P as primary elements.

Conceptually, the Minkowski sum is a dilation or expansion of the primary image B, whose form is determined by the generating polygon A. In the previous example we have chosen a circle as the generating image. The expansion of the primary image is directionally uniform or isotropic, since the generating image is a symmetrical figure with regard to both axes. Most of the times, the complexity is in the choose of the center coordinates within the buffers. Normally, an aleatory point is chosen to become the center of the first buffer. Each point is then assigned to the nearest center, and each center is recalculated as the mean of all points assigned to it. In most cases, some not necessary calculations are done in [5] proposing an algorithm based on left and right side partial derivatives that allow to accelerate the method to choose the centers of the new buffers.

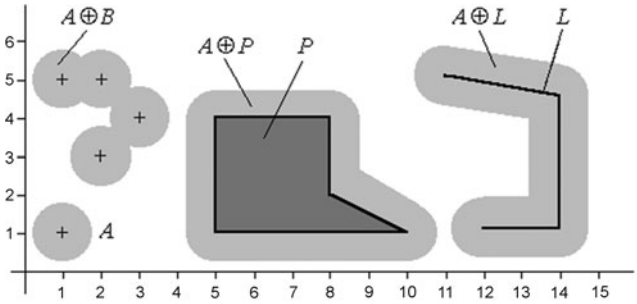

Fig. 6 Examples of the Minkowski sum taking a circle as the generating geometric object

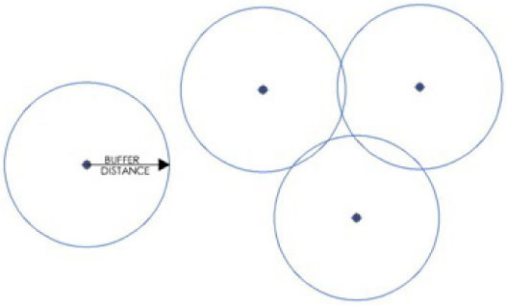

Fig. 7 Areas of influence of specific data with a fixed distance

4 Buffers in the Commercial GIS

The majority of commercial GIS have one or various modules that generate areas of isotropic influence. There are three methods to generate the buffer or the zone of influence: the creation of areas of influence of specific data, the creation of areas of influence of linear data and the creation of areas of influence of polygonal data. This study will centre on the creation of areas of influence of specific data.

4.1 The Creation of Areas of Influence of Specific Data

Very frequently, GIS requires the generation of areas of influence in certain operations that analyze spatial data. The simplest area of influence is generated around specific data, as the process only implies the creation of "circular" polygon around each point, with a radius that is equivalent to the distance from the buffer. There are two ways of assigning the width of the buffer, the first applies a fixed buffer distance (specified by the user) to all points of a layer (Fig. 7).

The second (Fig. 8) consists of assigning an individual distance to each point based on the attributes of another layer

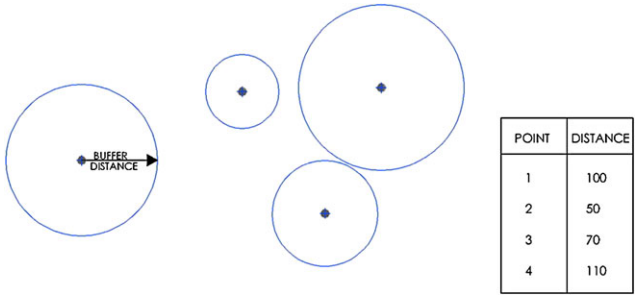

Fig. 8 Areas of influence of specific data with layer attributes

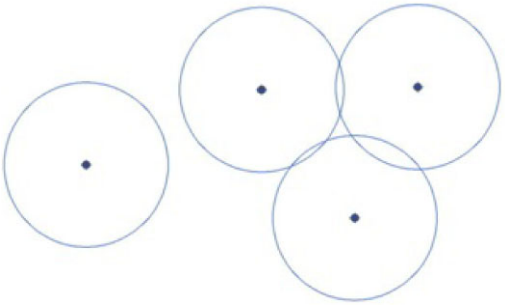

Fig. 9 Calculation of the intersection of areas of influence

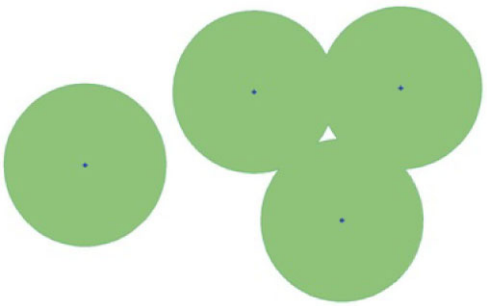

Fig. 10 Calculation of the dissolution of areas of influence of specific data

of the system (its weighting or weight). The width attributes of the area of influence in this case will be stored in a feature attribute table.

In case multiple points need to be analyzed in the layer, the system should test the existence of overlaps between the areas of each point (Fig. 9). These overlaps should be deleted in such a way that the result is a polygonal layer that represents the zone covered by the union between all the areas. Hence, this procedure implies the application of two additional operations: intersection and dissolution (Fig. 10).

The creation of buffers gives as a result a new poly-gonal-type layer in the system, which represents the areas of influence generated from both weighted and fixed value distances. The resulting table of polygons will contain the identifiers of the polygons created in the procedure, and a new attribute that indicates whether the polygon is found inside

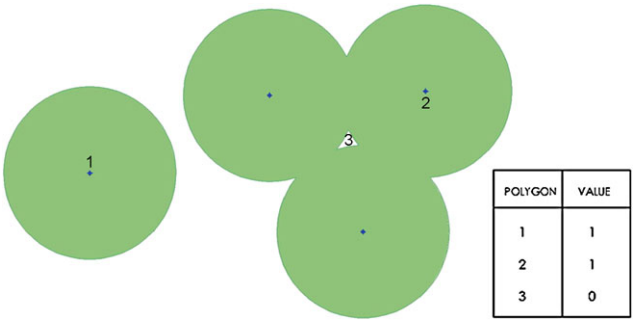

Fig. 11 Calculation of the dissolution of areas of influence of specific data with layer attributes

or outside the area of influence (Fig. 11). In this case it is necessary to model those areas that are not adapted to the circular shape. In [16] an algorithm is shown, used to simulate those kind of shapes, taking into account not only the point within the circle, but also those that are tangent to the polygon that circumscribes them.

5 A Forecasting Hybrid ISOTROPIC-CBR System

The images to be analyzed in this CBR system are divided into smaller squares. A squared zone determines the area that will be independently analyzed. The values of the different variables in a square area at a certain moment, which define the problem or the situation that has to be solved, is referred to as *a case*.

5.1 Case Base Creation

In this study, we have applied isotropic image analysis based on the buffer operator using Von Misses distribution and the Minkowski Sum, both previously introduced in Sect. 3. Owing to its good adaptation capabilities, this system has been applied to calculate the areas of different environmental phenomenon, so that they may be modeled.

Once the data is structured, it is stored in the *case base*. The temporal situation of each case, which relates it with the next situation in the same position is stored. That temporal relationship is what creates the union between *problem* and *solution*. The problem is the past case, and the solution is the future case, the future state of the square under analysis.

Growing Cell Structures (GCS) [21] are used when introducing the data into the case base. GCS can create a model from a situation organizing the different cases by their similarity. If a 2D representation is chosen to explain this technique, the most similar cells (*cases* in CBR) are near one or the other. If there is a relationship between the cells, they are grouped together, and this grouping characteristic helps the CBR system to retrieve similar cases in the next phase. When a new cell is introduced into the structure, the closest

cells move towards the new one, changing the overall structure of the system as shown in (11) and (12). The weights of the winning cell, ω_c, and its neighbours, ω_n, are changed. The new value is represented by $\omega_c(t+1)$, and $\omega_n(t+1)$ respectively. The terms ϵ_c and ϵ_n represent the learning rates for the winner and its neighbours, while x represents the value of the input vector.

$$\omega_c(t+1) = \omega_c(t) + \epsilon_c(x - \omega_c) \tag{11}$$

$$\omega_n(t+1) = \omega_n(t) + \epsilon_n(x - \omega_n) \tag{12}$$

5.2 Generating Predictions

Once the historical data is stored in the case base, and the GCS have been restructured according to the original distribution of the variables, the system is ready to receive a new problem. When a new problem is introduced into the system, GCS are used once again. The stored GCS behave as if the new problem were stored in the structure, and finds the cells (cases in the CBR system) that are the most similar to the problem introduced into the system. In this case, the GCS does not change its structure, because it is being used to retrieve the most similar cases to the introduced problem. Only in the retain phase does the GCS change, introducing the proposed solution once again, if it is correct.

The similarity of the new problem to the stored cases is determined by the GCS calculating the distance between them. Every element in the GCS has a series of values and the distance between the elements is therefore a multidimensional distance, where all the elements are considered to establish the distance between cells. Then, after obtaining the most similar cases from the case base, they are used in the next phase. The selected case bases will be used to generate an accurate prediction according to the previous solutions that relate to the problem that was introduced.

Once the most similar cases to the problem to be solved are recovered from the case base, they are used to generate the solution. The prediction of the future probability of finding oil slicks in an area is generated by using an artificial neural network with a hybrid learning system. An adaptation of *Radial Basis Functions Networks* is used to obtain that prediction [23, 34]. Basis Function networks were chosen because of their reduced training time, in comparison to other artificial neural network systems, such as *Multilayer Perceptron* (MLP).

Growing RBF networks [26] are used to obtain the predicted future values that correspond to the proposed problem. This adaptation of the RBF networks allows the system to grow during training, gradually increasing the number of elements (prototypes) which act as the centers of the radial basis functions. In this case the creation of the Growing RBF must be made automatically, which implies an adaptation of the original GRBF system [43]. The definition of the error for every pattern is shown below in (13):

$$e_i = l/p \sum_{k=1}^{p} \|t_{ik} - y_{ik}\| \tag{13}$$

where t_{ik} is the desired value of the kth output unit of the ith training pattern, y_{ik} the actual values of the kth output unit of the ith training pattern. The Growing RBF pseudocode is shown in Algorithm 1:

Algorithm 1 `Growing Radial Basis Function pseudocode`

```
1. Calculate the error, e_i (3) for
every new possible prototype.
   a. If the new candidate is not among
   those selected and the calculated error is
   less than a threshold error, then the new
   candidate is added to the set of accepted
   prototypes.
   b. If the new candidate already belongs to
   the accepted ones and the error is less
   than the threshold error, then modify the
   weights of the units, in order to adapt
   them to the new situation.
2. Select the best prototypes from the
candidates.
   a. If there are valid candidates, create a
   new cell centered on the valid candidate.
   b. Otherwise, increase the iteration
   factor. If the iteration factor reaches
   10% of the training population, freeze
   the process.
3. Calculate global error and update the
   weights.
   a. If the results are satisfactory, end
   the process.
   b. If not, go back to step 1.
```

Once the GRBF network is created, it is used to generate the solution to the proposed problem. The network is trained with the same set of historical data that is included in the case base in every moment. Training data is also used in the prediction generation process. The data used to train the GRBF network is stored as part of the case base and it can be used to generate future predictions without accumulating any additional "noise" to the prediction process.

The GRBF network stays under a training process that continues until the results are considered good enough. To determine when a result is good enough, historical cases not used in the training process are presented to the network and the values that the GRBF yields as outputs are compared with their corresponding historical values. The results are

considered valid if these "*Slick Area*" values differ within a threshold that is defined with respect to the real values. The threshold is calculated by taking account of the number of cases stored in the case base: the more information available, the better the solutions should be, so the threshold used will be lower. For example, when, within the system, the number of stored cases is under 1000 different elements, the difference between the predicted value and the actual one can be up to a 30% of the actual value and considered a valid solution. When the amount of information stored is increased to, for example, 3000 elements, the difference between the prediction and the actual measure should be under 15% to be considered a good prediction.

After the training process, the GRBF is used to generate the solutions using the cases retrieved from the case base. This solution consists of a number representing the area of sea covered with oil (in km^2). When a problem is introduced into the system, the network generates a solution for every recovered case similar to the problem and the average of those solutions is the proposed solution. The solution will be the output of the network using the selected cases from the case base as input data. When new data is introduced in the case base, the GRBF is trained again, to adapt it to the new elements that are introduced.

The correction of the proposed solution is known when using test data, as all test data is part of the historical information obtained at the time of the oil spill or after it, by acquiring it from different means (satellites, direct observations, etc.) so it is possible to compare the solution proposed by the system with the real registered value.

5.3 Revising the Proposed Solution

Once generated, the prediction shows the analyzed area divided into small squares to the user. The squares are colored depending on the presence or otherwise of slicks in those squares. The intensity of the color corresponds to the possibility of finding oil slicks in that area. The areas colored with a higher intensity are those with the highest probability of finding oil slicks in them. Representing the prediction by coloring the different small squared areas depending on the probability of finding oil slicks on them allows the user to check the correction of the proposed solution, comparing the proposed prediction with the actual data. But the system provides an automatic revision method that must, in either case, be checked by an expert user.

Explanations are used to check the correction of the proposed solution and to justify the solution [47]. To obtain a justification for a given solution, the cases selected from the case base are used once again. To create an explanation, a comparison between different possibilities was used. All the selected cases have their own associated future situation. If we consider the case and its solution as two vectors, we can establish a distance between them, calculating the evolution of the situation under the considered conditions. If the distance between the proposed problem and the solution given is not greater than the distances obtained from the selected cases, then the solution is a good one, according to the structure of the case base.

The explanations pseudocode is shown in Algorithm 2:

Algorithm 2 `Explanations pseudocode`

```
1. For every selected case in the retrieval
   phase, the distance between the case and
   its solution is calculated.
2. The distance between the proposed problem
   and the proposed solution is also
   calculated.
3. If the difference between the distance of
   the proposed solution and those of the
   selected cases is below a certain
   threshold value, then the solution is
   considered valid.
4. If not, the user is informed and the
   process goes back to the retrieval
   phase, where new cases are selected
   from the case base.
5. If, after a series of iterations, the
   system does not produce a good enough
   solution, then the user is asked to
   consider accepting the best of the
   generated solutions.
```

The distances are calculated by considering the sign of the values, not using its absolute value. This decision is easily justified by the fact that is not the same to move north as it is to move south, even if the distance between two points were equal. If the prediction is considered correct, it will be stored in the case base, and it can then be used in future predictions to obtain new solutions. It will have the same category as the historical data previously stored in the system.

When inserting a new case in the case base, GCS are used again. When adapting to the new solution introduced in the case base, the stored structure grows and improves its capability of generating good results since new knowledge has been introduced in the system. After explaining the system presented in this research, then, the results obtained by applying it to the oil spill problem will be shown. In next section, a resume of the results obtained with the presented system will be explained, as well as a comparison with previous solutions generated for the oil spill problem.

6 Application and Results

The novel system explained above has been applied to a real life case study to check the correction of the system itself

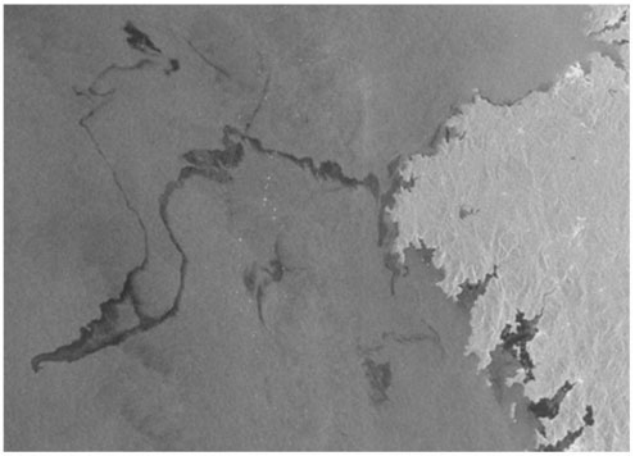

Fig. 12 Satellite SAR image of an oil spill near the north–western coast of Spain

as well as the isotropic image analysis. The system has been used to generate predictions in oil spill situations.

6.1 The Oil Spill Problem

After an oil spill issue, it is necessary to determine whether an area will be contaminated. To determine the presence or absence of contamination in an area, the behavior of the slicks generated by the spill has to be understood. The system presented here was trained using historical data acquired during the Prestige oil spill on the Galician west coast of Spain, from November 2002 to April 2003. Most of the data used by CROS was acquired from the ECCO (Estimating the Circulation and Climate of the Ocean) consortium (Menemenlis et al. [37]).

First of all, the position, shape and size of the oil slicks must be identified. The most precise way to acquire that information is by using satellite imagery. Synthetic Aperture Radar(SAR) images are the most commonly used to automatically detect this kind of slick [46]. These images have been interpreted using CBR systems both for monitoring [30] and for classification [11] purposes. The satellite images show certain areas where little or no activity is apparent, such as zones with no waves, that are in fact oil slicks. Figure 12 shows a SAR image of part of the western coastline of Galicia, along with some black areas corresponding to the oil slicks. With SAR images it is possible to distinguish between normal sea variability and slicks.

Once the slicks are identified, it is also crucial to know the meteorological and maritime conditions affecting the slick at the time of the analysis. Information gathered from weather satellites is used to obtain the required atmospheric data. That is how different variables such as temperature, sea height and salinity are measured in order to obtain a global model [48] that will explain how the slick is expected to evolve.

Table 1 Improvement in the results obtained after applying the buffer operator

Cases	RBF	O.CBR	GRBF+CBR	Isot.-CBR
100	1.2%	1.4%	1.7%	2.1%
500	2.7%	2.6%	2.9%	2.9%
1000	3.1%	3.2%	3.6%	3.8%
2000	3.7%	3.9%	4.4%	4.5%
3000	4.2%	4.6%	4.8%	5.4%
4000	4.6%	5.0%	5.2%	6.0%
5000	5.1%	5.3%	5.6%	7.2%

There have been different ways to analyze, evaluate and predict situations after an oil spill. One approach is by simulation [6], where a model of a certain area is created, introducing specific parameters (weather, currents and wind) and working with a forecasting system. Using this methodology, it is easy to obtain a good solution for a certain area [15], but it is quite difficult to generalize in order to solve the same problem in new zones. It is also possible to create a model for a specific and problematic area [41], which is of great, albeit limited assistance, because it is not possible to apply that same solution to different geographical areas. Current data must be considered in order to create contingency plans that could help to minimize environmental risks [8]. The end use of all these systems is in decision-support systems that can help to take all the decisions that need to be taken in an organized manner. To achieve this objective, different techniques have been used, ranging from fuzzy logic [32] to negotiation with multi-agent systems [31].

6.2 Results

Previous versions of the novel system presented here have been used to generate predictions to this very problem [35]. In this study, the use of the buffer operator introduces an improvement in the quality of the information. The available data is more accurate when using the buffer operator as a novel tool, which is key to an improved image analysis; the data stored in the case base, which is used to generate the solutions is more truthful, and so too are the predictions generated by the system: the more accurate and precise the information, the greater the improvement to the results.

Table 1 shows a summary of the results. Four different techniques were compared, using an incremental case base containing between 100 to 5,000 cases. The first technique, represented as "*RBF*" represents a simple Radial Basis Function Network that is trained with all the available data. The network receives an area and its parameters as an input. The RBF network gives a probability of finding oil slicks in the analyzed area, as an output, which is considered a solution to the problem. The second one, an "*Oceanic*

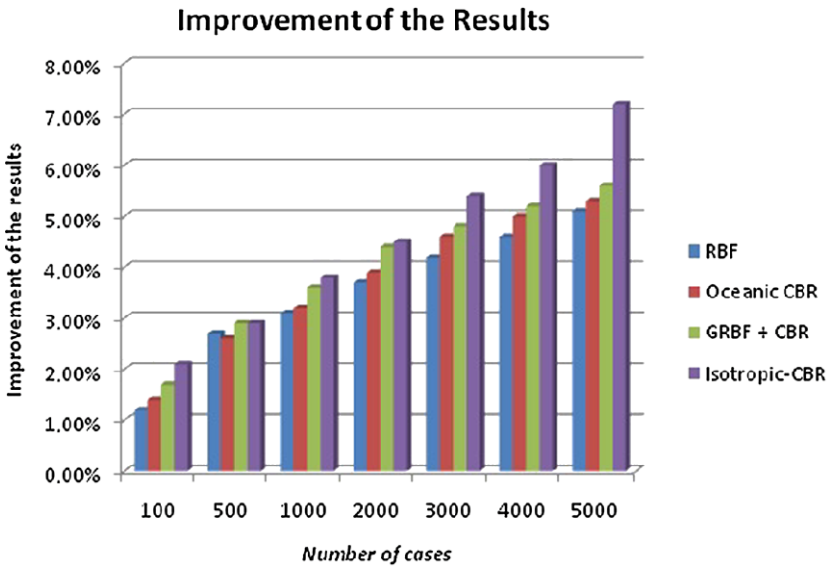

Fig. 13 Improvement of the results after applying the buffer operator

CBR" system represents a CBR system applied to forecast oceanographic methods [9]. This system uses neural networks in the adaptation process of the recovered cases, in particular a Radial Basis Function network. The neural network has a process of recovering elements from a network knowledge base, from where the neural network retrieves the parameters to calibrate the network. This CBR system has been applied to oceanographic problems. The Isotropic-CBR system presented here uses the GCS algorithm to structure the case base, improving the organizational characteristics of the case base and includes GRBF networks, which generate more accurate predictions than the RBF network. The third one, called "*GRBF+CBR*", corresponds to the possibility of using a GRBF neural network combined with CBR. Recovery from the CBR is carried out by using the Manhattan distance to determine the closest cases to the problem. The GRBF network works in the reuse phase, adapting the selected cases to obtain the new solution. Finally, the "*Isotropic-CBR*" system represents the system presented here. An increasing number of cases is due both to the analysis of additional images of oil spills added to the set of images used in previous tests and to the reuse of solutions proposed by the system. Table 1 shows the evolution of the results along with the increase in the number of cases stored in the case base. The numerical results showed in Table 1 represent the average of a series of tests completed with the available information. The number of tests performed in each iteration (every time the case base grows) represents ten percent of the size of the case base (if the case base contains 1000 elements, then 100 tests will be performed, and so on). When amount of available information is increased by adding new data to the system (new satellite data, new direct observations), the results are validated again considering the new conditions offered by the new data. The more data available, the better results the systems usually generate. The elements that make part of the ten percent of the data used to test the correction of the system has not previously been used in its training. Those cases also come to form part of the historical information and may be used to check the correction of the predictions generated by the system. For every case stored in the case base there is a future situation corresponding to the solution of that situation. The cases used to test the system are randomly chosen from the overall amount of cases. The results for each of the techniques under analysis improved when the number of cases stored was increased. Figure 13 shows a graphical representation of the results in Table 1, clearly showing that the use of the buffer operator in the Isotropic-CBR leads to an improvement in the results. Table 1 shows the improvement obtained in the different systems previously explained after applying the buffer operator to them. Those improvements are measured with a series of identical systems that only differs from the ones used here in the use of the buffer operator. It is quite significant to notice that all of them improve their results, what means that the quality of the information used to generate the solutions is better when using the buffer operator.

Table 2 Results obtained with the different techniques used for comparison

Cases	RBF	O.CBR	GRBF+CBR	Isot.-CBR
100	46%	41%	44%	45%
1000	53%	50%	61%	70%
3000	64%	62%	73%	86%
5000	68%	69%	77%	94%

Table 2 shows the average values of the accuracy of the predictions generated by the different systems used for com-

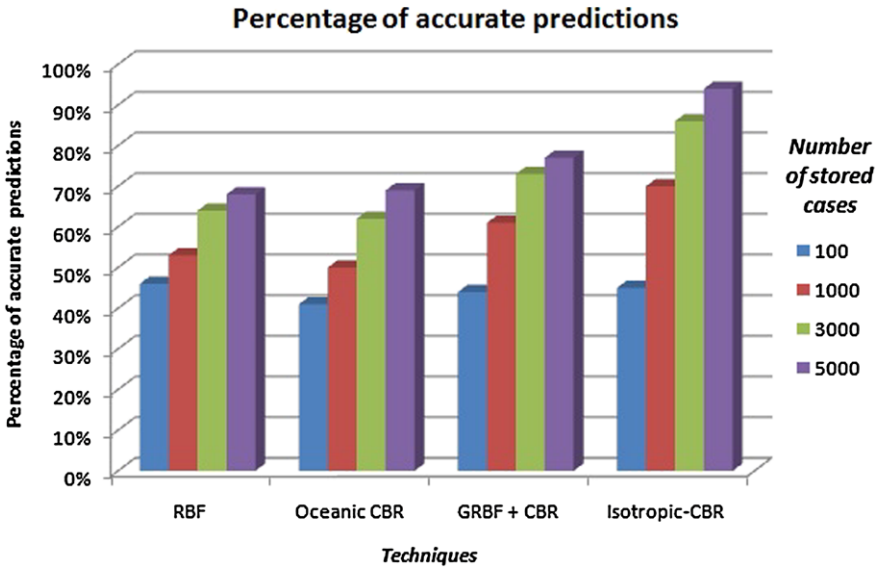

Fig. 14 Percentage of good predictions from the four different techniques and for different numbers of cases

parison after applying the buffer operator. The percentage of good predictions obtained by the different systems is shown in Table 2. It may be seen that while the number of cases stored in the system increases, the accuracy of the results also improves. The results generated by the Isotropic-CBR system are better than those of the other systems, especially when a large enough amount of information is stored in the system. Figure 14 shows a graphical representation of the data in Table 2. There is a group of columns for every technique analyzed. The different colours of the columns in the figure represent the growth of the case base, having different number of stored cases. It clearly shows the improvements obtained by the Isotropic-CBR system. When the case base has 100 elements, it can be seen how the four methods performs in a similar way. Nevertheless, and as expected, when the number of cases increases, the difference in performance is higher. When the case base contains 5000 cases, the percentage of good predictions achieve a quite high value for the novel hybrid model(Isotropic-CBR: 94%) when the second best model (GBRF+CBR) achieve a value of 77%. The improvement obtained when the amount of stored elements is increased is bigger in the presented model than in the other techniques. This is mainly due to the importance of the reuse of the information and the optimization of the use of past successful results.

7 Conclusions

We have presented a novel hybrid CBR system, by using, for the first time, a GIS technique based on the use of an isotropic buffer operator.

The areas in our CBR system were calculated by dividing the global images into smaller ones, so that a different buffer may be applied to each one. Changing the size of the buffer will help the system to generate a more accurate analysis, improving the quality of the data in the final case-based solution, resulting in better prediction results.

The system presented in this study has been applied to generate predictions in an oil spill environment. The results shown in Sect. 6 demonstrate the accuracy of the image analysis performed with the aid of the buffer operator. The "*Isotropic-CBR*" system presented here has been compared with three other systems, explained in the previous section, showing better results than those systems. When the amount of information available is increased, the results obtained by the "*Isotropic-CBR*" system are significant better than those obtained with the other systems. When having 5000 elements in the case base, the results obtained by the system presented here are between 17 and 24% better than those obtained with the other systems.

The next steps in the development of the buffer operator applied to this CBR system will be the application of the system to other case studies and improvements to image analysis and the use of the buffer operator; thereby introducing novel techniques that may generate better results in this and other fields of knowledge.

Acknowledgements This research has been partially supported through project of the Spanish Ministry of Science and Innovation TIN2010-21272-C02-01 (funded by the European Regional Development Fund).

References

1. Aamodt, A.: A Knowledge-Intensive, Integrated Approach to Problem Solving and Sustained Learning, Knowledge Engineering and Image Processing Group. University of Trondheim, Trondheim (1991)

2. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun. **7**(1), 39–59 (1994)
3. Althoff, K.D., Mänz, J., Nick, M.: Maintaining experience to learn: case studies on case-based reasoning and experience factory. In: Proc. 6th Workshop Days of the German Computer Science Society (GI) on Learning, Knowledge, and Adaptivity (LWA 2005). Saarland University, Germany (2005)
4. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: What's next? Artif. Intell. Med. **36**(2), 127–135 (2006)
5. Bretin, E., Lachaud, J.O., Oudet, É.: Regularization of discrete contour by Willmore energy. J. Math. Imaging Vis. **40**, 214–229 (2011)
6. Brovchenko, I., Kuschan, A., Maderich, V., Zheleznyak, M.: The modeling system for simulation of the oil spills in the Black Sea. In: 3rd EuroGOOS Conference: Building the European Capacity in Operational Oceanography, p. 192 (2002)
7. Carrascosa, C., Bajo, J., Julian, V., Corchado, J.M., et al.: Hybrid multi-agent architecture as a real-time problem-solving model. Expert Syst. Appl. **34**(1), 2–17 (2007)
8. Copeland, G., Thiam-Yew, W.: Current data assimilation modelling for oil spill contingency planning. Environ. Model. Softw. **21**(2), 142–155 (2006)
9. Corchado, J.M., Aiken, J.: Hybrid artificial intelligence methods in oceanographic forecasting models. IEEE SMC Trans. **32**(4), 307–313 (2002)
10. Corchado, J.M., Bajo, J., Abraham, A.: GERAmI: improving the delivery of health care. IEEE Intell. Syst. **3**(2), 19–25 (2008) Special Issue on Ambient Intelligence
11. Chen, F., Wang, C., Zhang, H., Zhang, B., et al.: SAR images classification using case-based reasoning method. In: Geoscience and Remote Sensing Symposium, IGARSS 2007, pp. 2048–2051 (2007)
12. Chou, Y.H.: Exploring Spatial Analysis in Geographic Information Systems. Onward Press, Santa Fe (1997)
13. Decker, B., Rech, J., Althoff, K.D., Klotz, A., et al.: eParticipative process learning–process-oriented experience management and conflict solving. Data Knowl. Eng. **52**(1), 5–31 (2005)
14. Diaz, F., Fdez-Riverola, F., Corchado, J.M.: Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray data sets. Comput. Intell. **22**(3/4), 254–268 (2006)
15. Elhakeem, A.A., Elshorbagy, W., Chebbi, R.: Oil spill simulation and validation in the Arabian (Persian) Gulf with special reference to the UAE coast. Water, Air, & Soil Pollution. Focus **184**(1), 243–254 (2007)
16. Emre, M.: Improving the performance of k-means for color quantization. Image Vis. Comput. **29**, 260–271 (2011)
17. Fdez-Riverola, F., Corchado, J.M.: FSfRT: forecasting system for red tides. Appl. Intell. **21**(3), 251–264 (2004)
18. Fdez-Riverola, F., Iglesias, E.L., Díaz, F., Méndez, J.R., et al.: Applying lazy learning algorithms to tackle concept drift in spam filtering. Expert Syst. Appl. **33**(1), 36–48 (2007)
19. Fisher, R.A.: Dispersion on a sphere. Proc. R. Soc. London **217**, 295–305 (1953)
20. Fritzke, B.: Unsupervised clustering with growing cell structures, Neural Networks, 1991. In: IJCNN-91-Seattle International Joint Conference, p. 2 (1991)
21. Fritzke, B.: Growing cell structures-a self-organizing network for unsupervised and supervised learning. Neural Netw. **7**(9), 1441–1460 (1994)
22. Goodman, J.E., O'Rourke, J.: Handbook of Discrete and Computational Geometry. CRC Press, Boca Raton (2004)
23. Haykin, S.: Neural Networks. Prentice Hall, Upper Saddle River (1999)
24. Herrero, Á., Corchado, E., Pellicer, M.A., Abraham, A.: MOVIH-IDS: a mobile-visualization hybrid intrusion detection system. Neurocomputing **72**(13–15), 2775–2784 (2009)
25. Hsu, C.-C., Ho, C.-S.: A new hybrid case-based architecture for medical diagnosis. Inf. Sci. **166**(1–4), 231–247 (2004)
26. Karayiannis, N.B., Mi, G.W.: Growing radial basis neural networks: merging supervised andunsupervised learning with network growth techniques. IEEE Trans. Neural Netw. **8**(6), 1492–1506 (1997)
27. Kolodner, J.L.: Case-Based Reasoning. Kaufmann, Los Altos (1993)
28. Ladrón de Guevara, I., Muñoz, J., de Cózar, O.D. Blázquez, E.B.: Robust fitting of circle arcs. J. Math. Imaging Vis. **40**, 147–161 (2011)
29. Lee, J.H., Ha, S.H.: Recognizing yield patterns through hybrid applications of machine learning techniques. Inf. Sci. **179**(6), 844–850 (2009)
30. Li, X., Yeh, A.G.: Multitemporal SAR images for monitoring cultivation systems using case-based reasoning. Remote Sens. Environ. **90**(4), 524–534 (2004)
31. Liu, X., Wirtz, K.W.: Sequential negotiation in multiagent systems for oil spill response decision-making. Mar. Pollut. Bull. **50**(4), 469–474 (2005)
32. Liu, X., Wirtz, K.W.: Decision making of oil spill contingency options with fuzzy comprehensive evaluation. Water Resour. Manag. **21**(4), 663–676 (2007)
33. Mardia, K.V., Jupp , P.E.: Directional Statistics. Wiley, New York (2000)
34. Martín, B., Sanz, A.: Redes neuronales y sistemas borrosos. Editorial Ra-Ma, Zaragoza (1997)
35. Mata, A., Corchado, J.M.: Forecasting the probability of finding oil slicks using a CBR system. Expert Syst. Appl. **36**(4), 8239–8246 (2009)
36. Mena, M.: Aplicaciones de estadística circular a problemas de ciencias naturales. Akadia, Seftigen (2004)
37. Menemenlis, D., Hill, C., Adcroft, A., Campin, J.M., et al.: NASA supercomputer improves prospects for ocean climate research. EOS Trans. **86**(9), 89–95 (2005)
38. Mitra, S., Hayashi, Y.: Neuro-fuzzy rule generation: survey in soft computing framework. IEEE Trans. Neural. Netw. **11**(3), 748–768 (2000)
39. Montani, S., Portinale, L., Leonardi, G., Bellazzi, R.: Case-based retrieval to support the treatment of end stage renal failure patients. Artif. Intell. Med. **37**(1), 31–42 (2006)
40. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial tessellations: Concepts and applications of Voronoi diagrams (POD), Eur. Syst. Automat. 43 (2009)
41. Pavón, R., Díaz, F., Laza, R., Luzón, V.: Automatic parameter tuning with a Bayesian case-based reasoning system. A case of study. Expert Syst. Appl. **36**(2), 3407–3420 (2008)
42. Periáñez, R., Pascual-Granged, A.: Modelling surface radioactive, chemical and oil spills in the Strait of Gibraltar. Comput. Geosci. **34**(2), 163–180 (2008)
43. Plaza, E., McGinty, L.: Distributed case-based reasoning. Knowl. Eng. Rev. **20**(03), 261–265 (2006)
44. Ros, F., Pintore, M., Chrétien, J.R.: Automatic design of growing radial basis function neural networks based on neighboorhood concepts. Chemom. Intell. Lab. Syst. **87**(2), 231–240 (2007)
45. Schneider, P.J., Eberly, D.H.: Geometric Tools for Computer Graphics Geometric. Kaufmann, San Francisco (2002)
46. Schou, G.: Estimation of the concentration parameter in von Mises-Fisher distributions. Biometrika **65**(2), 369 (1978)
47. Solberg, A.H.S., Storvik, G., Solberg, R., Volden, E.: Automatic detection of oil spills in ERS SAR images. IEEE Trans. Geosci. Remote Sens. **37**(4), 1916–1924 (1999)
48. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning-perspectives and goals. Artif. Intell. Rev. **24**(2), 109–143 (2005)

49. Stammer, D., Wunsch, C., Giering, R., Eckert, C., et al.: Volume, heat, and freshwater transports of the global ocean circulation 1993–2000, estimated from a general circulation model constrained by World Ocean Circulation Experiment (WOCE) data. J. Geophys. Res. 108 (10.1029) (2003)
50. Tsai, C.Y., Chiu, C.C.: A case-based reasoning system for PCB principal process parameter identification. Expert Syst. Appl. **32**(4), 1183–1193 (2007)
51. Watson, I.: Case-based reasoning is a methodology not a technology. Knowl.-Based Syst. **12**(5–6), 303–308 (1999)
52. Wu, J., Yu, Y.: Connectionism-based CBR method for distribution short-term nodal load forecasting, TENCON 2005. IEEE Region **10**, 1–6 (2005)
53. Yang, B.S., Han, T., Kim, Y.S.: Integration of ART-Kohonen neural network and case-based reasoning for intelligent fault diagnosis. Expert Syst. Appl. **26**(3), 387–395 (2004)
54. Zhang, F., Ha, M.H., Wang, X.Z., Li, X.H.: Case adaptation using estimators of neural network. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, p. 4 (2004)