An analysis about the measure quality of similarity and its applications in machine learning

Yaima Filiberto[1] Rafael Bello[2] Yaile Caballero[1] Mabel Frias[1]

[1] Universidad de Camagüey, Camagüey, Cuba
[2]Universidad Central de Las Villas, Villa Clara, Cuba
{yaima.filiberto, yaile.caballero, mabel.frias}@reduc.edu.cu, rbellop@uclv.edu.cu

Abstract

In this paper, a review about the quality of the similarity measure and its applications in machine learning is presented. This measure is analyzed from the perspective of the granular computing. The granular computing allows analyzing the information at different levels of abstraction and from different approaches. The analysis shows that this measure is based on two basic aspects on the universe of objects: the granularity of the information and the principle that, similar problems have similar solutions. Using the measure, a method was formulated to build relations of similarity; these relations and other results have been used in improving machine learning techniques.

Keywords: quality of similarity measure, granular computing, rough set theory, machine learning

1. Introduction

The granular computing is a new computational term that includes the theories, methodologies, techniques and tools that make use of the granules (subsets of a universe) in the solution of problems, [1, 2]. The main concept is the granule; this is defined as a nucleus, entity or focal point of compound knowledge for different objects, mutually indistinguishable [3].

A granulation of the universe consists on decomposing the universe in a family of subsets (granules) that contains all the objects of the universe. The granulation depends on the relationship R. A universe U can be divided in a series of granules, each one settled down by a relationship R [4]. The combined quotient U/R contains groups of inseparable objects according to R. Two common types of granulation of the universe are the partition and the covering of the universe; in the first case, the intersection between granules is empty, in the second case the granules can be superimposed, and in both cases the union of all the groups is equal to the universe. The Rough Set Theory (RST) [5] and the Fuzzy Set Theory (FST) [6] offer concrete models of granular computing [7].

The granulation of the universe facilitates the learning process based on different perspectives of the information; it is possible to diminish the computational cost of the discovery process, as well as to find more abstract relationships. The RST is among the learning techniques that use the granulation. The rough sets have been used in the construction of algorithms for the rules generation, the feature selection, and other tasks of machine learning, like it is shown in [8, 9, 10, 11].

The classic RST is defined taking into account features with discrete domains; therefore, the methods for the knowledge discovery based on the classic RST are affected when mixed data (application domains in which the features can have discrete or continuous values) appears. To solve this problem, new concepts and methods have been developed based on similarity relations instead of equivalence relation; this is known as the extended Rough Set Theory [12]. The new problem is finding the appropriate similarity relation for each application domain. A method to build similarity relations based on the similarity quality measure is proposed in [13, 14], this includes to compute the weights for the features.

The effectiveness of the similarity relation and the weights for the features that are found using this approach has been proved when using these results for improving or creating the machine learning method. The purpose of this paper is to analyze all these data. The content of this paper is the following. In the section 2, the similarity quality measure is analyzed in the context of the granulation of the universe and the principle that "similar problems have similar solutions"; the method for building similarity relations is presented in section 3; in section 4, some improvements to techniques of machine learning are revised.

2. Analysis of the similarity quality measure.

The definition of different relationships among the objects determines different granulation alternatives. In the context of the supervised learning (classification problems or functions approximation problems), the knowledge that is discovered establishes relationships between the granulation regarding the condition features and the granulation regarding the decision fea-

ture (also called objective feature). Keeping in mind the relevance of this fact, measures have been defined that allow establishing a relationship grade between both granulations; for example, the quality of similarity measure in the RST; this measure has been employed in the development of machine learning techniques, [9, 15].

Let $DS = (U, A \cup \{d\})$ be a decision system, where A is the set of condition features and d is the decision feature. Let $B \subseteq A$ and $X \subseteq U$, B defines an equivalence relation and X is a concept. X can be approximated using only the information contained in B by constructing the B-lower and B-upper approximations of X, denoted by B_*X and B^*X respectively, where $B_*X = \{x \in U : [x]_B \subseteq X\}$ and $B^*X = \{x \in U : [x]_B \cap X \neq \varnothing\}$, and $[x]_B$ denotes the equivalence class of x according to B-indiscernible relation. The objects in B_*X are surely members of X, while the objects in B^*X are possibly members of X. Let a granulation $Y = \{Y_1, ..., Y_n\}$ of U according to the values of the decision feature d (classes), where Y_i, denote a decision class; in this case the set Y is a partition of U.

The coefficient $\gamma_B(Y)$ defined by expression (1) is called the quality of the classification of Y according to the features in B. It expresses the percentage of objects which can be correctly classified into classes $Y_1, ..., Y_n$ using the features in B only.

$$\gamma_B(Y) = \frac{\sum_{i=1}^{n} |B_*Y_i|}{|U|} \tag{1}$$

This is an important measure in the RST because it is related with important concepts of this theory, such as the consistency of the decision system, and the concept of reduct. A reduct is a minimal set of attributes $B \subseteq A$ such that $IND(B) = IND(A)$, $IND(X)$ is called the X-indiscernibility relation; this definition can be reformulated using the quality of the classification measure in the following way: a reduct B is a minimal set of attributes such that $\gamma_A(Y) = \gamma_B(Y)$.

It is possible to appreciate that the classification quality measure establishes a relationship between the granulation generated by the equivalence relations B and the granulation generated by the classes in the decision system. The concepts of conditional granularity (CG) and the decision granularity (DG) have been introduced to call the granularity of universe according to the conditional features and the decision feature respectively; in [16] the relationships between CG and DG are studied; therefore, the classification quality measure establishes a grade of relation between CG and DG.

But the classification quality measure is limited to the case of decision feature with discrete domain, that is, the case of classification problems where DG is a partition of the universe according to the decision classes. To overcome this restriction the similarity quality measure is introduced in [17], this measure is defined according to a similarity relation, which allows working with mixed data to describe the objects and working with discrete or continuous decision features. This measure can be analyzed as a combination of the granularity of the universe according to CG and DG, and the principle of the Case-based reasoning.

The solution in case-based reasoning starts from exploiting the relationship between two types of similarities, one defined on the space of the objects' description (condition features) and other one defined on the space of the solutions (features of decision) [18] and [19]; starting from a correct description of the problems, similar problems should have similar solutions. The principle of case-based reasoning can be expressed as "while more similar are the problems according to the condition features, more similar should be the values of the decision features".

The solution of problems using the case-based reasoning, and in general the lazy learning, has as a base that this principle is satisfied; in other case, they obtained solutions without the required quality. The similarity quality measure represents the degree in which the similarity between objects using features of condition is the same as the similarity according to the decision feature; and it is built in the following way.

Let be relationships R_1 and R_2 which determine the CG and DG respectively. Granules N_1 and N_2 defined for (2) and (3) can be built using R_1 and R_2, N_1 and N_2 of x are the neighbourhoods of x according to the relations R_1 and R_2 respectively:

$$N_1(x) = \{y \in U : xR_1y\} \tag{2}$$

$$N_2(x) = \{y \in U : xR_2y\} \tag{3}$$

By using the Dice similarity coefficient [20] and the granules N_1 and N_2, the expression (4) allows to calculate a similarity grade between both granules:

$$\varphi(x) = \frac{|N_1(x) \cap N_2(x)|}{0.5 * |N_1(x)| + 0.5 * |N_2(x)|}$$

$$0 \leq \varphi(x) \leq 1 \tag{4}$$

Calculating (4) for all the objects of the universe, the grade in which the similarity according to the condition features coincides with the similarity according to the decision feature can be measured. This grade is called the similarity quality measure, denoted by $\theta(DS)$, and it is defined by expression (5):

$$\theta(DS) = \left\{ \frac{\sum_{\forall X \in U} \varphi(x)}{|U|} \right\} \tag{5}$$

This measure does not require that the features of the decision system have discrete domain, like it is the case of the classification quality measure; moreover, it can be used in domains with mixed data. This measure depends on the granulation of the universe according to the condition features and the granulation due to the decision feature. These granulations can be partitions or coverings; they are defined by relations R_1 and R_2. In the next section, a method to build the relations is proposed.

3. Method to build similarity relations using the similarity quality measure

One of the most important advantages of the RST to develop problem solving techniques is that it practically does not require any additional information, only those contained in the universe of objects. Against this statement, it could be only mentioned that the classic formulation of the RST supposes the domain of the features is discrete, because it uses an indiscernibility relation of the data which defines as inseparable those objects that have same values for the subset of features considered in the relation ($xRy \Leftrightarrow bi(x) = bi(y)$ for all $bi \in B \subseteq A$, and $x, y \in U$). Defined in this way, this relation is an equivalence relation.

As it was established before, this type of indiscernibility relation imposes a restriction for the case of domains with mixed data. For example, in the case of features with continuous domains, slight differences among objects are not meaningful when building the equivalence classes, where the measurement mistakes of the attribute's values always lead to an imprecise description of the objects we are working on. The answer to this problem has been discretizing the continuous domains or to work with other alternatives of indiscernibility relation; see figure 1; the first alternative reduces the problem to the classic RST and the second one leads to the extend RST. A generalization of the classical rough set approach is specified by the replacement of the equivalence relation with a binary, weaker similarity relation. This yields some extensions to the classical RST such as [12, 21, 22, 23, 24, 25]. The purpose is to extend the inseparability relation R so as to gather into the same class those objects which are not identical but closer (similar) enough to a referent object according to the similarity relation. While an equivalence relation defines granules like equivalence classes, the similarity relation defines the granule as a similarity classes. The similarity class of x, according to the similarity relation R is denoted by $R(x)$ and defined by (6):

$$R(x) = \{y \in U : yRx\} \tag{6}$$

It is read as "the set of elements in U that are similar to according to R".

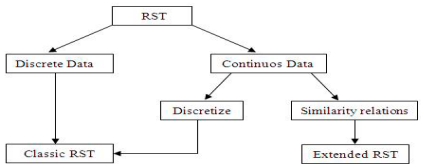

Fig. 1: Alternatives in RST.

The equivalence relations induce partitions of the universe U, while the similarity relations induce a covering of a universe. A covering of universe U is a family of subsets of U where no subsets in the family are empty and the union is equal to U. A partition of U is a covering of U, so the concept of a covering is an extension of the concept of a partition. Many researches have been developed to study covering approximation space in the RST, such as [25, 26, 27]; some of them oriented to build a covering-based generalized RST.

However, the determination of the most appropriate similarity relation for an application domain does not have an immediate answer; becoming an outstanding problem in order to use the methods based on the extended RST. Subsequently, it is analyzed the solution proposed in [13, 14] to build similarity relations. The similarity relation is formulated as $xRy \Leftrightarrow F(x, y) \geq \varepsilon$, x, $y \in U$, and $F(x, y)$ is a similarity function with values in [0,1].

The function F includes the following terms, and it can be defined by expression (7):

- Local similarity measures used to compare the values of single features (called comparison functions of the feature).
- Feature weights representing the relative importance of each attribute.
- A global similarity measure responsible for the computation of a final similarity value based on the local similarities and feature weights (called similarity function).

$$F(x, y) = \sum_{i=1}^{n} w_i * \partial_i(x_i, y_i) \tag{7}$$

where: n is the number of features w_i is the weight of feature i X_i and Y_i are the values of feature i in objects X and Y respectively ∂_i is the comparison function of feature i.

Given the features comparison function's ∂_i (an example is showed in expression (10)), the problem of building the relation R is reduced to find the set of weights associated to the features, that is $W = \{w_1, w_2, ..., w_n\}$.

Considering the granulation of the universe according to the condition's features and the decision's feature (conditional granulation and decision granulation), the relations R_1 and R_2 are defined according to (8) and (9):

For all objects x and y in U:

$$xR_1y \; if \; and \; only \; if \; F_1(x,y) \geq \varepsilon_1 \quad (8)$$

$$xR_2y \; if \; and \; only \; if \; F_2(x,y) \geq \varepsilon_2 \quad (9)$$

Where F_1 and F_2 are similarity functions to compare objects in U, F_1 includes features in A and F_2 computes the similarity degree between two objects according to the value of the decision feature d; ε_1 and ε_2 are thresholds. If F_1 is defined according to the expression (7), the problem is to find the set of weight W that maximizes the similarity quality measure (expression (5)).

This is an optimization problem that can be solved using a heuristic search, in which the heuristic evaluation function defined by the similarity quality measure is maximized. In [13, 14], the Particle Swarm Optimization (PSO) [28] has been used to solve the problem. In this paper (to see section 4.1) an extension of the study is presented by using as heuristic method the Univariant Marginals Distribution Algorimth (UMDA), instead of PSO. The UMDAc algorithm, proposed by [29], has interesting characteristics such as the relative ease of implementation, speed in locating the optimal solution, its powerful exploring capabilities and its relative lower computational cost in terms of memory and time. The method to calculate the set of weights, that use a heuristic search method (as PSO or UMDA) to maximize the similarity quality measure (expression (5)) is called maxQS.

4. Applications in Machine Learning

The quality similarity measure and the method to build similarity relations described in previous sections allow making a granulation of the universe and extending the RST in the case of mixed data. The quality of this granulation and this similarity relation has been shown in several applications in the machine learning field. In this section, three cases are revised that show the applicability of the similarity quality measure and the derived results from it. The weights for the condition features have been used to build the initial set of weights for a multilayer net, which improves the learning algorithm performance. The similarity function defined for the expression (7) and the weights for the condition features have been used to retrieve similar instances in the k-NN method, and the similarity relation was used to propose a new method for prototype-based learning; the effectiveness of both was proved. Moreover, an algorithm has also been formulated to discover classification rules in domains with mixed data.

4.1. Application in MLP.

The Artificial Neuronal Network called Multilayer Perceptron (MLP) is a powerful model for solving nonlinear problems. The construction of neuronal nets is a problem of non lineal optimization in which the objective is to find a set of weights that minimizes the cost function. This cost function is generally characterized by a great number of local minima in the vicinities of the global minimum.

The network topology and the initial weights play a very important role. In general, the formation of the MLP network is carried out through the successive intents with different network topologies and sets of weights until arriving to satisfactory results for the problem. The learning process includes an initialization algorithm and an algorithm of weights adjustment. The weight can be initialized in a random way, but it is important to keep in mind that the results will depend in great measure of the value of these weights [30],that is an interesting problem as it is shown in [31, 32, 33].

In general, the multilayer Perceptron can have several hidden layers. However, in the study presented in [14, 34], the initialization of the MLP with a single hidden layer is considered. The values for the weights of the connections between the input nodes and the nodes in the hidden layer are initialized using the weights of features according to the maxQS method described in section 3. The weights for the connections between the nodes of the hidden layer and the nodes in the output layer are generated randomly.

In [14], a study of the performance of the MLP is presented with this variant of initialization of the weights in the problem of function approximation. The problem is to calculate the resistant capacity of three types of connectors (stud, crestbond and canals); it is an important parameter because it is responsible of ensuring the connection among structures. Three databases were used; the output for each instance in the three databases is the value of resistant capacity. Two initialization alternatives were studied: the first one appears in the implementation of the MLP in the Weka tool [1], where the weights are initialized in a random way; the other is initializing the weights using the weight according to the method maxQS. The results obtained by the MLP with both initialization variants were compared with the real value of the resistant capacity, using three different error measures (Mean Absolute Percentage Error, Root Mean Square Error, and the average magnitude of the difference between the desired value and that obtained by the prediction); in all the cases the obtained error by the new method of initialization of the weights was smaller.

[1]Tool of open code written in Java. Available under GNU public licenses in http://www.cs.waikato.ac.nz/Ÿml/weka/

$$\partial(x_i, y_i) = \begin{cases} 1 - \frac{|x_i - y_i|}{Max(a_i) - Min(a_i)} & if\ i\ is\ continuous \\ 1 & if\ i\ is\ discrete\ and\ x_i = y_i \\ 0 & if\ i\ is\ discrete\ and\ x_i \neq y_i \end{cases} \quad (10)$$

On the other hand, in [34] an expanded study of this new method of weight initialization is presented in classification problems and the function approximation problems. There were used 24 databases from the UCI repository [2], 12 where the domain of decision attributes is nominal (classification) and 12 where is numerical (function approximation). The performance of the MLP using four different types of weights was studied: random generation, calculation of the weights by the conjugate gradient (method KNN_{VSM} [35]), using the same weight value for all attributes ($w = 1/numAtt$), and the weights calculated by the method maxQS. The error measures used are Mean Absolute Percentage Error and the Average of the differences between the desired and produced value by the method. In the classification problems, weights calculated by using the Relieff method (RELIEF) [36] is used also. In all the cases the performance of the MLP using the weight calculated according to the maxQS method was superior to the rest. These studies show that the calculated weights according to the maxQS method, described in section 3, are better than other sets of weights when they are used to initialize the neuronal network MLP, in both case, using training sets coming from real problems or international databases as the UCI Repository.

In this paper a new experimental study of the performance of the maxQS method is presented, in which the search method UMDA is used instead of PSO.

The problem is to calculate the resistant capacity for the channel type connector. The dataset has five input variables and one output with a total of 43 instances. The input variables are: thickness of the soul (w), thickness of the wing (t), longitude of the connector (L), height of the connector (H), resistance of the concrete to the compression (fc). The output for each instance is the value of resistant capacity (Q). In the experimentation, the expression (10) has been used as comparison function in the expression (7), which allows working with mixed data. A three-layer neural network with inputs features, outputs, and one hidden layer with variable number of nodes $(n+q)/2$ was designed.

The result obtained using this set W in the MLP method was compared with other four alternatives to W. The k-fold cross-validation process was employed. K - Fold Cross - Validation divides the original dataset into k subsets of equal size where one is used as test set while the others as used as the training set. Then the overall accuracy of the classifier is calculated as the average precision obtained with all test subsets. This technique eliminates the problem of overlapping test sets and makes an effective use of all available data. The recommended value $k = 10$ was used [37].

Five alternatives of methods to calculate the weights were employed for the experimentation. The variants for calculating the weights are: (i) the method maxQs using UMDA (called UMDAc+RST), (ii) the method maxQs using PSO (called PSO+RST), (iii) assigning the same weight to each feature (called Standard), (iv) Random (MLP-R) and (v) calculation of the weights by the conjugate gradient method (KNN_{VSM}).

The results obtained were compared with the real value of the resistant capacity according to the measures: (i) Mean Absolute Percentage Error (MAPE) and (ii) Mean Absolute Error (MAE). In order to compare the results, we will use a multiple comparison test to find the best algorithm. In Table 1 can be observed that the best ranking is obtained by UMDAC+RST for the MAE measure. Iman_Davenport test [38] is carried out (employing F-distribution with 4 and 12 degrees of freedom) in order to find statistical differences among the algorithms, obtaining a p-value near to zero. In this manner, Table 2 shows the results of the Holm [39] procedure for comparing UMDAc+RST to the remaining ones. The algorithms are ordered with respect to the obtained z-value. Thus, by using the normal distribution, the corresponding p-value associated with each comparison can be obtained and this can be compared with the associated α in the same row of the table to show whether the associated hypothesis of equal behavior is rejected in favor of the best ranking algorithm (UMDAc+RST), the test rejects three cases. It can be noticed that the method based on UMDA is statistically superior to Standard, MLP-R and KNN_{VSM} methods, and it presents comparable results with PSO+RST. In order to process

Table 1: Results of the Friedman statistical test's for the MAE error measure.

Algorithm	**Ranking**
UMDAc+RST	1
PSO+RST	2.25
Standard	4.5
MLP-R	2.75
KNN_{VSM}	4.5

[2]UCI Machine Learning Repository. In C. U. o. C. Irvine. (Ed.). http://www.ics.33.edu/ mlearn/ MLRepository.html

Table 2: Holm test α=0.5, taking as method of control UMDA+RST.

i	**algorithm**	$z = (R_0 - R_i)/SE$	p	**Holm**	**Hypothesis**
4	Standard	3.130495	0.001745	0.0125	reject
3	KNN_{VSM}	3.130495	0.001745	0.016667	reject
2	MLP-R	1.565248	0.117525	0.025	reject
1	PSO+RST	1.118034	0.263552	0.05	accepts

Table 3: Summary of the comparison of the method of UMDAc+RST with the different calculation methods.

Parameters	$Qexp/$					
	CSA	$NRMC$	$SENACYT$	$Pachan$	$PSO+RST$	$UMDAc+RST$
Arithm. Mean	2.4024	3.0065	0.9881	1.6151	0.9987	1.0001
Max Value	35.95	28.72	67.83	54.99	73.56	69.87
Min Value	5.98	5.98	19.59	9.35	21.17	21.25
Stand. Deviation	15.61	16.57	13.08	15.12	13.05	12.53
Correl. Coeff.	0.96	0.96	0.98	0.74	0.98	0.99
$0.85 \leq Qexp/Q \leq 1.15$	0	0	41	7	42	43
$Qexp/Q < 0.85$	0	0	1	2	1	0
$Qexp/Q > 1.15$	43	43	1	1	34	0

the results of the experiments we used KEEL [40].

The best results are obtained when PSO+RST and UMDAc+RST, next they are compared with other classical methods for the prediction of the resistant capacity of connectors (Q). These methods are: CSA [41], NRMC [42], SENACYT [43] and Pashan [44]. These results show that the most stable methods are UMDAc+RST and PSO+RST. Table 3 show these results.

This experimental results suggest that the construction of similarity relations based on similarity functions (such as the defined by (7) and the similarity quality measure defined by (5)) is feasible, independently of the heuristic method that is used.

4.2. Applications in lazy learning.

4.2.1. Effect on the performance of the method k-NN.

The k-nearest neighbor's method (k-NN) [45] is one of the well-known and relatively simple methods to solve classification and functions' approximation problems.

The basic idea of the k-NN method is that similar input values have similar output [19]. The output of an object is calculated starting from the output from the most similar neighbors to it; for example, in the case of the classification, the k-NN rule classifies each unlabeled example by the majority label of its k-nearest neighbors in the training set. To select the similar instances' set it is necessary to use a function that allows calculating the similarity grade between two objects. In the classic k-NN method, the Euclidean distance is used to find the nearest neighbors because the domains of all features are the real numbers.

Different studies have shown that the performance of k-NN depends crucially on the way that distances are computed between different examples, among them [46, 47]. The results presented in [48] show that an important aspect in the methods based on similarity grades, as the k-NN method, is the set of weights assigned to the features, because this improves significantly the performance of the method.

The importance of building the most appropriate similarity measure in the case-based systems, particularly to solve function approximation problems, is studied in [49]. This is the so-called problem of distance metric learning. An alternative solution to this problem consists in using a similarity function like the defined by the expression (7) setting as the weight features those calculated by the maxQS method described in the section 3, which maximize the similarity quality measure.

The employment of this type of similarity function is studied in [13] to give solution to the problem of the prediction of the resistant capacity of stud connectors in composite structures using the k-NN method like functions approximation. The following alternatives were studied to calculate the weight: assigning the same weight to each feature, three alternatives based in the expert criteria and the weights calculated by the maxQS method. To calculate the error the following measures are used: MAPE and MAE. The experimental results show that when the last alternative is used to calculate the weights, the error of the approximation is significantly lower; there are significant differences in accuracy with respect to the real value. In [14], this study is extended to other two types of con-

nectors (crestbond and canals) using the alternatives mentioned before, plus the weights obtained by Conjugated Gradient method (KNN_{VSM}) [35], and the three error measures; the experimental results also indicate that the weighted similarity function using the weights that maximize the similarity quality measure allows to minimize the errors.

In this paper, the study shown in [13, 14] is extended by using 12 data set's for classification problems (tae, bridges_version1, diabetes, biomed, iris, zoo, schizo, soybean-small, cars, heart-statlog, liver-disorders and glass) and 12 data sets for functions' approximation problems (baskball, bodyfat, detroit, diabetes_numeric, elusange, fishcatch, pollution, pw-linear, pyrim, sleep, vineyard and scholvote) from UCI Repository, eight of them are mixed datasets.

In the functions approximation problems three alternatives of weights were compared for the similarity function (Standard (1/Quantity-features), KNN_{VSM} and maxQS), to evaluate the effectiveness of the k-NN method were also used the two error measures, before mentioned. For the statistical analysis of the results the techniques of hypothesis test were used [50]. For multiple comparisons, the Friedman tests are used and of Imam-Davenport test [38] to detect differences statistically significant among a group of results. The Holm test [39] was used also. Tables 4 and 5 show the result (Imam-Davemport (distribution of F with 2 and 22 grades of freedom) value of p: 0.000016789863). The KEEL tool's in its version 2.0 was used in the statistical analysis in all experimental results; specifically the Non-Parametric Statistical Analysis module was use [40].

The Holm test was applied, it shows the results of the measures using the weights computed by the maxQS method are significantly lower to those obtained when the weights are calculated by means of the Standard methods and KNN_{VSM} methods. Friedman's test and Imam-Davemport's test were applied among the Standard method, KNN_{VSM} method and maxQS method, regarding the correlation coefficient of Pearson, to the R2 coefficient and MAE measure and it is shown that significant differences exist among them.

A similar experimental study was carried out for the case of 12 classification data sets. In this case, the Relief method for the weight calculation is included also. A Friedman's test and Imam-Davemport's test were applied, among the Standard methods, KNN_{VSM} methods, Relief methods and maxQS methods regarding the general accuracy of the classification of the k-NN method; it is demonstrated that significant differences exist among them. The Holm's test was applied, regarding the general accuracy of the classification, and it is corroborated that the general accuracy of the classification is significantly superior when the weights are calculated by the maxQS method that when they are calculated by the Standard methods, KNN_{VSM} methods and Relief methods. This study of the accuracy ratifies that k-NN method produces the best results when the similarity function defined (7) joint to the weights calculated using the maxQS method is used.

4.2.2. A method for building prototypes using the similarity relation.

In the Nearest Prototype Learning (NPL) [51], the idea is to determine the value of the decision feature of a new object analyzing his similarity with regard to a set of prototypes, selected or generated from an initial set of instances. The prototypes represent the typical characteristics of a set of instances instead of necessary or sufficient conditions; the prototypes can be abstractions of the same instances previously observed, or they can be the directly observed examples. To learn prototypes is to represent the information of the training sets as a set of points in the space of the application domain, called prototypes, the decision value of a new point is calculated using the decision value of one or more prototypes. The intention of the NPL is to decrease the costs of storage and processing of the learning techniques based on instances.

In [52], the NP-BASIR method is proposed to construct prototypes using similarity relations. The performance of the method was studied using the similarity relations based on the similarity quality measure in the function approximation problem. For each instance in the training set, the similarity class is built using the similarity relation; a prototype is build using an aggregation operator over the instances in this similarity class. The intention is to construct a prototype or centroid for a set of similar objects. The performance of the NP-BASIR algorithm was studied using 19 data sets of the UCI's Repository. It is possible to appreciate that the proposed method achieves a substantial reduction of the quantity of instances (in the majority of the cases a reduction about to 80 per cent of the number of instances), preserving the precision. This result is very important due to the computational complexity of the lazy methods (such as k-NN and nearest prototype) and it depends on the quantity of instances; when the number of instances decreases, the computational cost is reduced. Also, that paper shows, that there are not significant differences between the efficacy of the approximation of functions obtained using the set of prototypes and the rule of the most similar neighbors and other aproximators like Multilayer Perceptron, Linear Regression and Regression Tree.

Table 4: Results of the Friedman statistical test's for the MAE error measure.

Algorithms	Ranking
Standard	2.4167
KNN_{VSM}	2.5
PSO+RST	1.0833

Table 5: Holm test's for $\alpha = 0.5$ for the MAE error measure, taking as control PSO+RST.

i	Algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Hypothesis
2	KNN_{VSM}	3.47011	0.00052	0.025	rejects
1	Standard	3.265986	0.001091	0.05	rejects

4.3. IRBASIR: an algorithm for learning classification rules.

Using the similarity relation built according to the method described in section 3, the IRBASIR algorithm (Induction of Rules BAsed on Similarity Relations) for the induction of classification rules was formulated; it allows to discover knowledge using decision systems with mixed data without the necessity of carrying out a discretization process before or during the knowledge discovery process.

The rules represent functions that establish a relation between the examples (described by means of a set of features) and the decision classes. They are usually expressed in the way $if\ P\ then\ Q$, where P is the conditional part formed by a conjunction of elementary conditions ($p_1\ and\ p_2\ and....p_k$), and Q is the decision part that assigns a value of decision (class) to an object that completes the condition.

However, the IRBASIR algorithm has a different way to express P. The algorithm induces rules in the way $if\ \sum w(i) * \partial_i() \geq \varepsilon\ then\ Q$, where $w(i)$ is the feature weight i, $\partial_i()$ is the comparison function for the feature i and ε it is a threshold. The algorithm performs an iterative procedure in which each instance in the training set, not used before, is employed to generate a decision rule; the similarity class of the instance is built using the similarity relation, and a decision rule is generated in the way $if\ \sum w(i) * \partial_i() \geq \varepsilon\ then\ Q$, where Q is the value of the majority class in the similarity class. The weights and the similarity relation are calculated according to the method described in the section 3.

In [53], the behavior of the IRBASIR algorithm is studied using 12 datasets of the UCI Repository with mixed data. In the experimentation, the results obtained by IRBASIR were compared with two versions of the C4.5 algorithm (the C4.5 classifiers in the KEEL tool [40], and J48 in the Weka tool), and the MODLEM algorithm in the ROSE2 tool [3]. In practically all cases the accuracy of the classification reached using the rules induced by the IRBASIR algorithm was superior to those generated by the C4.5 algorithm, J48 algorithm and MODLEM algorithm. Multiple comparison tests were used to compare the results with the purpose of finding the best algorithm, the statistical analysis confirms that the IRBASIR algorithm is statistically better than others.

[3]Rough Sets Data Explorer. Available in http://www-idss.cs.put.poznan.pl/rose

5. Conclusions

In this paper a review has been presented with the following results: the similarity quality measure; the calculus of the features weights by means of the optimization of this measure, the construction of the similarity relation based on weighted similarity functions that include those weights, and the employment of previous results in improving machine learning techniques. The study includes the analysis of the quality of similarity measure from the perspective of the granular computing and the principle of the Case-based reasoning. It is also studied a method to build similarity relations based on weighted similarity functions, which consists in maximizing the value of this measure by using a metaheuristic.

A revision of the effectiveness of the derived results from the measure is presented in the context of the machine learning techniques. This includes a new analysis about the performance of the k-NN method, enlarging the previous study, using trainings sets of the UCI's Repository.

References

[1] Y. Y. Yao. The art of granular computing. page 7, 2007.

[2] W. Pedrycz, A. Skowron, and V. Kreinovich. *Handbook of Granular Computing*. Wiley, Chichester, West Sussex, 2008.

[3] L. Zadeh. Is there a need for fuzzy logic? *Information Sciences*, 178:2751–2779, 2008.

[4] J. M. Ma and et al. Granular computing and dual galois connection. *Information Sciences*, 177:5365–5377, 2007.

[5] Z. Pawlak. Rough sets. *International Journal of Information & Computer Sciences 11*, 11:145–172, 1982.

[6] L. A. Zadeh. Fuzzy sets. *Information and Control 8*, pages 338–353, 1965.

[7] Y. Y. Yao. Probabilistic approaches to rough sets. *Expert Systems with Applications*, 20(5):287–297, 2003.

[8] J.-W. Grzymala-Busse. A new version of the rule induction system lers. *Fundamenta Informaticae*, 31:27–39, 1997.

[9] Y. Gomez and et al. Two step swarm intelligence to solve the feature selection problem. *Journal of Universal Computer Science*, 14(15):2582–2596, 2008.

[10] K. Thangavel and A. Pethalakshmi. Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing*, 9:1–12, 2009.

[11] L. James and et al. A dominance-based rough set approach to customer behavior in the airline market. *Information Sciences*, 180:2230–2238, 2010.

[12] Z. Pawlak and A. Skowron. Rough sets: Some extensions. *Information Sciences*, 177:28–40, 2007.

[13] Y. Filiberto, R. Bello, Y. Caballero, and R. Larrua. Using pso and rst to predict the resistant capacity of connections in composite structures. In *In International Workshop on Nature Inspired Cooperative Strategies for Optimization Springer*, pages 359–370. Berlin: Springer-Verlag, 2010.

[14] Y. Filiberto, R. Bello, Y. Caballero, and R. Larrua. A method to built similarity relations into extended rough set theory. In *IProceedings of the 10th International Conference on Intelligent Systems Design and Applications ISDA2010 IEEE*, pages 1314–1319. IEEE Press, 2010.

[15] X. Wang and et al. Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters 28*, pages 459–471, 2007.

[16] B. Chen and et al. Granular rough theory: A representation semantics oriented theory of roughness. *Applied Soft Computing*, 9:786–805, 2009.

[17] Y. Filiberto, R. Bello, Y. Caballero, and R. Larrua. Using pso and rst to predict the resistant capacity of connections in composite structures. *In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO, Springer, Heidelberg*, 284:359–370, 2010.

[18] D. B. Leake. Cbr in context: The present and future. in case-based reasoning: Experiences, lessons, and future directions. *Menlo Park: AAAI Press/MIT Press*, 1996.

[19] R.L. Lopez and E. Armengol. Machine learning from examples: Inductive and lazy methods. *Data & Knowlege Engineering 25*, 25:99–123, 1998.

[20] E. Deza and M. Deza. *Dictionary of distances*. Elsevier, 2006.

[21] R. Slowinski and D. Vanderpooten. Similarity relation as a basis for rough approximations. in wang, p.p. (ed) advances in machine intelligence & soft-computing. *Duke University Press, Durham, NC*, IV:17–33, 1997.

[22] Y.Y. Yao. Relational interpretations of neighborhood operators and rough set approximations operators. *Information Sciences 101*, pages 239–259, 1998.

[23] S. Greco. Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129:1–47, 2001.

[24] A Skowron and et al. Approximations spaces and information granulation. *Lectures Notes on Computer Science 3400*, pages 175–189, 2005.

[25] K. Qin and et al. On covering rough sets. *Lectures Notes on Artificial Intelligence 4481*, pages 34–41, 2007.

[26] W. Zhu and F. Wang. Relations among three types of covering rough sets. pages 43–48. In IEEE GrC. Atlanta, USA., 2006.

[27] D. Bianucci. Entropies and co-entropies of coverings with applications to incomplete information systems. *Fundamenta Informaticae*, 77:77–105, 2007.

[28] J. Kennedy and R.C. Eberhart. Particle swarm optimization. In *Proc. IEEE International Conference on Neural Networks*, pages 1942–1948, 1995.

[29] P. Larrañaga, R. Etxeberria, J. A. Lozano, and J. M. Peña. Optimization by learning and simulation of bayesian and gaussian networks. Kzzaik-4-99, Dept. of Computer Science and Artificial Intelligence, University of the Basque Country, 1999.

[30] M. Faúndez-Zanuy. Nonlinear speech processing: Overview and possibilities in speech coding. *In: Chollet, G., Esposito, A., Faúndez-Zanuy, M., Marinaro, M. (eds.) Nonlinear Speech Modeling and Applications. LNCS (LNAI). Springer, Heidelberg.*, 3445:15–42, 2005.

[31] L. M. Almeida and T. B. Ludermir. An evolutionary approach for tuning artificial neural network parameters. *LNAI*, 5271:156–163, 2008.

[32] P. Nieminen and T. Kärkkäinen. Ideas about a regularized mlp classifier by means of weight decay stepping. *In: Kolehmainen, M., Toivanen, P., Beliczynski, B. (eds.) ICANNGA 2009. LNCS. Springer, Heidelberg*, 5495:32–41, 2009.

[33] X. Fu and et al. A resource limited immune approach for evolving architecture and weights of multilayer neural network. *LNCS*, 6145:328–337, 2010.

[34] Y. Filiberto, R. Bello, Y. Caballero, and G. Ramos. Improving the mlp learning by using a method to calculate the initial weights of the

network based on the quality of similarity measure. *I. Batyrshin and G. Sidorov (Eds.): MICAI 2011, Part II, LNAI, Springer-Verlag Berlin Heidelberg*, 7095:351–362, 2011.

[35] D. Wettschereckd. A description of the mutual information approach and the variable similarity metric. In *Technical report, Artificial Intelligence Research Division*, Sankt Augustin, Germany, 1995. German National Research Center for Computer Science.

[36] I. Kononenko. Estimating attributes: Analysis and extensions of relief. In *Proc. European Conf. on Machine Learning*, pages 171–182, 1994.

[37] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[38] R.-L. Iman and J.M. Davenport. Approximations of the critical region of the friedman statistic. *Commu. Stat.*, 18:571–595, 1980.

[39] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[40] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 2010.

[41] CSA. *Handbook of steel constructions. (8va ed.)*. Canadian Institute of Steel Construction (2001). Canada: [s.n.], 2001.

[42] NRMC-080-2007. *Calculation of between floors made up of concrete and steel with soul beams full subjected to load static. Code of good practical 2007: Brunch Norma of the Ministry of the Construction of Cuba*, 2007.

[43] O. Ramírez, R. Larrúa, R. Vargas, F. Yeomans, and M. Pinto. Caracterización experimental detallada de conectores en estructuras compuestas de hormigón y acero.proyecto senacyt col 006-007. fundamentación experimental de sistemas estructurales y productos para el desarrollo competitivo de la construcción compuesta. Technical report, 2010.

[44] A. Pashan. Behaviour of channel shears connectors: pus-outs tests. Master's thesis, University of Saskatchewan, Saskatchewan, Canada., 2006.

[45] T.-M. Cover and P.-E. Hart. Nearest neighbour pattern classification. volume 13, pages 21–27. Institute of Electronical and Electronics Engineers Transactions on Information Theory, 1967.

[46] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. *Neighbourhood components analysis.*, volume 17. Advances in Neural Information Processing Systems, Cambridge, MA., 2005.

[47] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[48] W. Duch and K. Grudzinski. Weighting and selection features. intelligent information system viii. In *Proceedings of the Workshop held in Ustron*, pages 32–36, Ustron, Poland, June 14-18 1999.

[49] A. Stahl. Learning feature weights from case order feedback. In *Lecture Notes in Computer Science Springer*, volume 2080, pages 502–516, 2001.

[50] S. García and et al. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180:2044–2064, 2010.

[51] C. James, J. Bezdek, and L. Kuncheva. Nearest prototype classifier designs: An experimental study. *International journal of Intelligent systems*, pages 1445–1473, 2001.

[52] M. Bello, M. Garcia, and R. Bello. A method for building prototypes in the nearest prototype approach based on similarity relations for problems of approximation of functions. 2012.

[53] Y. Filiberto, R. Bello, and Y. Caballeroand M. Frias. Algorithm to learn clasification rules based on the extended rough set theory. *DYNA año 78*, 169:62–70, 2011.