

A Context Similarity Measure

Yoram Biberman

Department of Mathematics and Computer Science
Ben-Gurion University of the Negev
P.O.B. 653, 84105 Beer-Sheva, Israel
e-mail: yoramb@cs.bgu.ac.il

Abstract. This paper concentrates upon similarity between objects described by vectors of nominal features. It proposes non-metric measures for evaluating the similarity between: (a) two identical values in a feature, (b) two different values in a feature, (c) two objects. The paper suggests that similarity is dependent upon the context: It is influenced by the given set of objects, and the concept under discussion. The proposed Context-Similarity measure was tested, and the paper presents comparisons with other measures. The comparisons suggest that compared to other measures, the Context-Similarity suites best for *natural concepts*.

1 Introduction

The notion of similarity is fundamental in many areas of cognition and computer science. The most frequent approach to similarity is to interpret it as a closeness in a spatial sense. According to this approach, objects are represented as points in a geometrical space; similarity between objects is inversely related to the distance between the objects, where the distance is measured by some metric function.

Many similarity measures exist in the literature (cf. [6] for a short survey of different measures), yet most of them are truly suited for continuous or ordered variables, but not for nominal (symbolic, unordered) ones. The most frequently used similarity measure for nominal domains is the City-Block one. The City-Block measure evaluates the similarity between every two identical values in a feature as equals to one, and the similarity between every two different values as equals zero. The similarity between two objects is the sum of the similarities of their features.

This paper concentrates upon similarity in the context of learning. ExeMplar-Based LEarning Models (EMBLEMs) suggest that concepts are learned by memorizing examples; the main information the learner needs to store in his/its memory is the classified examples the teachers supply; no general information in the form of rules is induced; the learner classifies new examples by comparing them to stored exemplars. EMBLEMs are represented in machine learning by Aha Kibler and Albert's *Instance-Based Learning (IBL)* [2], Stanfill and Waltz's *Memory-Based Reasoning* [11], Salzberg's *Generalized Exemplar Theory* [9], *Protos* of Porter et al. [7], and others. These models are based upon similarity between objects, yet, as described above, for nominal domain variables some of

them use the naive City-Block or modifications of it as their similarity measure. Aha Kibler and Albert write: “we have not yet experimented with sophisticated definitions for defining similarity for symbolic-valued attributes.” ([2], p. 62)

Following Tversky [12], this paper suggests that similarity is not a fixed property, but is dependent upon the context, that is, *the given set of examples*, and *the concept under discussion*. For example, if we present subjects the countries: USSR, China, Germany, Austria, they would tend to say that USSR is most similar to China, but if we replace Austria by Taiwan most subject would indicate that USSR is most similar to Germany. This finding evidences that the similarity between a pair of examples is dependent upon other examples that are presented, i.e., upon the context. Similarly, the similarity between USSR and Cuba is perceived as large when the concept under discussion is political system, but as small when the relevant concept is geographical area. Thus the similarity between two objects is not fixed for all contexts, but changes. Tversky notes that “judgments [of] similarity depend on context and frame of reference that, in turn, are determined by the nature of the task and the set of objects under consideration.” ([12], p. 340)

The next section presents my motivation and intuition. Sections 3 and 4 present functions for evaluating similarity between: (a) two primitive values in a feature (e.g. the colors red and green), (b) two examples (e.g. an apple and an orange). Section 5 presents experiments with the the proposed Context-Similarity measure, and suggests that among the measures that were tried Stanfill and Waltz’s Value Difference Metric (VDM) [11] performs best on domains that contain many irrelevant features, while the Context-Similarity performs best with *natural concepts*; Sect. 5.1 explains which concept are considered *natural*. Briefly stated, natural concepts can not be defined by necessary and sufficient conditions, are structured, may change with time and context, and generally have no sharp boundaries.

2 Motivation

The City-Block measure evaluates two different values in a nominal scale feature as contributing zero to the overall similarity between their examples. In many cases this evaluation is inaccurate as even in nominal domains humans differentiate between different degrees of similarity (e.g. the color red is more similar to orange than to blue, the material paper is more similar to textile than to metal, the faculty computer science is more similar to mathematics than to literature). Medin and Schaffer [5] suggest that the similarity between every pair of nominal values should be represented by a unique parameter. They do not specify how this parameter would be set. Protos [7], an EBL system, implements an elaborated similarity measure for nominal variables; in Protos this is achieved using the assistance of an expert that supplies the necessary domain theory (e.g. the colors red and green are equivalent for the purpose of classifying apples). Here I would like to propose a few principles that may guide one in an effort to evaluate the similarity between two different nominal values. The aim is to evaluate this

similarity in a more subtle manner than assuming it is zero, without relying on the assumption that this knowledge is supplied by an external source.

Metric similarity measures assume that the contribution of two equal values to the perceived similarity between their examples is equal for all values. I also question this assumption: Think of two people who are anarchists versus two people that hold main-stream political views; or think of two left handed people as opposed to two right handed people. I suggest that two examples that share an exceptional value (property) in some feature are perceived as more similar one another than two examples that share a common value. In other words, there is a ‘pop-out’ effect to a unique value which causes the examples to be perceived as similar to each other, and stand in contrast to the rest of the examples. Tversky [12] demonstrated this idea by saying that we perceive two identical twins as more similar to each other than two identical cars; the reason to this phenomenon is that there are only two identical twins ‘of a certain model’, but many identical cars of a certain model.

A third property of many of the existing measures is that their similarity evaluation is not affected by the concept under discussion. This assumption is also inaccurate in many cases; for example a concept like ‘apples’ contains both green and red members; thus red and green should be considered similar for the task of classifying apples; but cucumbers are always green, therefore if one needs to classify cucumbers she should consider red and green as dissimilar; another example that demonstrates this idea involved the ‘items’ USA and USSR; these two countries were similar with respect to the concept ‘the Great Powers’ but not with respect to ‘communist countries’. I, therefore, suggest that when a system needs to evaluate the similarity between two objects it should do so with respect to a certain concept.

Finally, following Medin and Schaffer [5], Hintzman [4] and Porter et al. [7] it is proposed that the similarity between two objects is determined *by an interaction* of the different components. This assumption is in contrast to independent cue models, like the City-Block one, which assume that the overall similarity is a function of the independent components that are summed.

These considerations lead me to propose a set of non-metric functions for evaluating similarity between examples described by nominal attributes, in the context of a categorization task.

3 Similarity of Features

This section proposes a similarity measure between two values in a feature. The next section utilizes this measure, and presents a measure of similarity between objects.

The features similarity measure considers two cases: whether the two values to be compared are equal or not.

Denote the similarity function between two values by s_{val} ; this function is composed of s_{eq} which computes the similarity between two equal values, and

$s_{d_i f}$ that handles the case where the values are different; finally S_{ex} denotes the examples similarity measure.

3.1 Elements with Equal Values

In contrast to the minimality axiom of the metric approach, I suggest that the similarity between two identical elements is not necessarily equal for all values (this implies that the triangle inequality also does not hold). The similarity between two examples that share a same value v in a feature is perceived as larger if v is less frequent in the population.

Denote by P the set of all examples supplied by the teachers, by $|P|$ the size of P , and by $|v|$ the number of examples with a value v . The contribution of two equal values $v_{1f} = v_{2f}$ in a feature f to the similarity between the examples that contain them, E_1 and E_2 , is negatively correlated with their frequency in P . Therefore it is defined to be: $s_{eq}(v_{1f}, v_{2f}) = \frac{|P| - |v_{1f}|}{|P|}$.

Yet, the similarity between the two values might be an artifact that contributes nothing to the categorization task, or might be irrelevant for the concepts that should be learned; but then the the occurrence rate of the value in the different concepts would be more or less equal (e.g., if being in favor of the Israeli-Palestinian agreement is meaningless for the categorization of members of the Israeli parliament into the different parties then the proportion of supporters of this agreement in the different parties would be more or less equal). In other words, the variance of the value's occurrence rate in the different concepts would be small. Denote this variance by $var(v, CS)$. We, therefore, conclude that for the task of categorization, the contribution of two equal elements to the similarity between their examples, should be defined as:

$$s_{eq}(v_{1f}, v_{2f}) = \frac{|P| - |v_{1f}|}{|P|} \cdot var(v_{1f}, CS) .$$

3.2 Elements with Non-Equal Values

The above expression evaluates the contribution of two equal values to the similarity between their examples. For a pair of non-equal values I suggest the following considerations:

- **Similarity between different values:** If two objects (e.g. an orange and a lemon) belong to the same concept, and share equal values in many features (e.g. texture, kind-of-peeling, juicyness, season-in-which-eaten), then we tend to perceive them as generally similar, and conclude, that in particular in those features where they have different values (e.g. oranges' color is orange while lemons are yellow) the values are similar with respect to the concept under discussion. In other words, the similarity between two different values v and u in a feature f with respect to a concept C , is positively influenced by pairs of examples that belong to C , contain the values v and u in f , and share equal values in many other features. In a dual manner:

- **Difference between different values:** If two objects (e.g. an apple and a tomato) belong to different concepts (e.g., one is a fruit while the other is a vegetable), and share equal values in many features (e.g., texture, kind-of-peeling, juiciness, taste), then these features make them relatively similar; in order to explain why they belong to contrasting concepts we tend to assume that v and u are relatively different, and they, at least partially, cause to the split between concepts. In other words, the similarity between two different values v and u in a feature f , with respect to a concept C , is negatively influenced by pairs of examples E_1, E_2 , where E_1 belongs to C , E_2 does not belong to C , E_1 has the value v (or u) in f , E_2 has the value u (or v) in f , and E_1, E_2 share many other properties.

An implementation of the above principles is as follows: Let E^v and E^u be two examples that share equal values v_1, \dots, v_m in a subset of their features, but have different values v and u in some feature. The effect of this pair over the perceived similarity between v and u can be expressed as $effect(E^v, E^u, v, u) = \sum_{i=1}^m s_{eq}(v_i, v_i)/n$, (where n is the total number of features used to describe each example). If both E^v, E^u belong to a same concept C then this effect is positive over $s_{dif}(v, u, C)$; if, on the other hand, E^v belongs to C , while E^u belongs to C' (or vice versa) then their effect over $s_{dif}(v, u, C)$ is negative. $s_{dif}(v, u, C)$ is calculated by averaging over all pairs of examples that contain v, u , and at least one of them belong to C . Formally expressed:

$$s_{dif}(v, u, C) = \frac{\sum_{E_v \in S^v, E_v \in C, E_u \in S^u, E_u \in C} effect(E_v, E_u, v, u)}{n_1} - \frac{\sum_{E_v \in S^v, E_v \notin C, E_u \in S^u, E_u \in C} effect(E_v, E_u, v, u)}{n_2} - \frac{\sum_{E_v \in S^v, E_v \in C, E_u \in S^u, E_u \notin C} effect(E_v, E_u, v, u)}{n_3}$$

where S^v denotes the set of examples that have a value v , n_1, n_2, n_3 denote the number of examples that enter into each of the three summations.

3.3 Elements with Missing Values

In many real life domains some feature values are missing from certain examples. In these cases the similarity measure should estimate the similarity between a missing value and present one, or between two missing values.

Different estimations can be made in such cases. Aha et al. [1] choose a ‘cautious’ approach: If either values of a pair is missing, then it is assumed that the two values are maximally different from each other. The Context-Similarity proposes a more optimistic estimation, namely the similarity between a value v and a missing value (with respect to a concept C) is estimated as the average similarity of v to every other value (with respect to C), weighted by the relative frequency of this value (e.g., if in some feature there are three different values v, u and w ; such that $s_{dif}(v, u, C) = 1$, $s_{dif}(v, w, C) = 0$, and there are twice

many examples that have u than w ; under these assumptions, the similarity between v and a missing value with respect to the concept C equals 0.66). The similarity between two missing values is evaluated along the same line as the average similarity between every pair of values with respect to C .

Formally expressed: Let v, v_1, \dots, v_m be the set of values in some feature, $s_{dif}(v, v_1, C), \dots, s_{dif}(v, v_m, C)$ the similarity between v and each v_i ($i = 1, \dots, m$) with respect to some concept C ; $|v|$ denotes the number of examples that have the value v . The similarity between v and a missing value $-$ is defined to be:

$$s_{dif}(v, -, C) = \frac{\sum_{i=1}^m s_{dif}(v, v_i, C) |v_i|}{|v_1| + \dots + |v_m|}$$

4 Similarity of Examples

The previous section described the contribution of two values in a feature to the perceived similarity between their examples; denote this similarity by $s_{val}(v_1, v_2)$.

The similarity between two examples $E_1 = (v_{11}, \dots, v_{1m})$, and $E_2 = (v_{21}, \dots, v_{2m})$ is defined to be:

$$S_{ex}(E_1, E_2) = \left(\sum_{i=1}^m s_{val}(v_{1i}, v_{2i}) / m \right)^r$$

where r is an odd natural number, larger than one.

The only point of interest in this definition is the condition over r . A larger value of r will yield larger interactions between the values of the different features, as it would produce larger terms, i.e., terms that are composed of more factors. An odd r is necessary in order to preserve the sign of the sum (that might be negative). This consideration was proposed by Medin and Schaffer [5] in their context model, and by Hintzman [4] in his MINERVA 2—learning from examples model.

Raising the sum by a power larger than one has another property: If a classifier that needs to classify an example E sums the similarity of E to a subset of the different concepts then raising the sum of the features similarity by a power greater than one ‘amplifies’ the similarity of more similar examples, and cuts down the similarity of less similar items, thus creating a preference to a set that contains few items very similar to E over a set that has more members that are less similar to E . For example, assume that the sums of the features similarities of E to two members of a concept C_1 are 0.5 and 0, while these sums to two members of C_2 are 0.25 and 0.25. If the sum of the feature similarities is not raised by a power, then E is equally similar to the two concepts; if, on the other hand, the sum of the features similarity is raised by a power greater than one, then E is found to be more similar to C_1 than to C_2 . In the experiments described below the value of r is irrelevant.

5 Experimental Results and Discussion

The presented similarity measure was tested on a variety of examples. The next subsection presents the domains that were used in the experiments. Section

5.2 presents two experiments that compare between the Context-Similarity and four other similarity measures: The City-Block measure, Stanfill and Waltz's [11] Weighted Feature Metric (WFM) and Value Difference Metric (VDM), and Tversky's Contrast similarity measure [12].

5.1 The Examples that were Used in the Experiments

This section overviews the domains that were used in the experiments, and their main properties. This overview would later enable us to draw preliminary conclusions regarding the question "which similarity measure is best suited for which class of concepts?". I suggest that this kind of research is no less important than proposing a new similarity measure, as no measure would be best for all domains, therefore it would be desired if we could match between the possible measures and the different domains.

The examples that were used were obtained from the repository of machine learning databases cited in the University of California, Irvine (UCI). Figure 1 presents a statistic overview of the examples.

Anderson and Matessa write: "It is informative to engage in horse races between learning algorithms. Different algorithms will work optimally given different data sets, and it should be our first task to understand the characteristics of the domains to which the algorithms are adapted" ([3], p. 293). Psychologists and philosophers distinguish between *natural concepts* versus *logical* ones. The idea is that logical concepts can be defined by a set of necessary and sufficient conditions, have sharp boundaries, and are unstructured—all members of a concept belong to the concept and represent it equally (prime numbers, grandmothers are examples of logical concepts). Natural concepts, on the other hand, can not be defined by a set of rules, have no sharp boundaries, and are structured—some members represent their concepts better than others. Most concepts humans use in everyday life are natural (e.g. furniture, vehicles). Three of the databases that are used in the following experiments are good examples of domains that contain natural concepts: (a) The 'Zoo' domain contains different kinds of animals. (b) The 'LED display' example can be described as an 'artificial natural concept': It is artificial on the one hand, as it is produced by a computer program, but it has many characteristics of natural concepts on the other hand; it also resembles the kind of concepts psychologists use in their laboratory experiments that aim to investigate human categorization in natural domains. (c) 'Hayes-Roth and Hayes-Roth (1977)' is an example of a database that was borrowed from such experiments. The characterization of other concepts with respect to the 'logical' versus 'natural' property is less obvious.

In some databases, the number of attributes used to describe each example is large. Two of these databases (#5 and #6) describe different diseases, and their attributes were proposed by experts in these fields. One can wonder about their dimensionality, and whether an expert, when diagnosing a patient, considers this number of variables. The answer is 'no'; some of the attributes are irrelevant for some of the diseases, and are not used by a human expert. The complete set is needed in order to cover all the possible categories, but every

The domains that were used in the experiments

Domain #	Domain	435	16	2	2	61	y
1	1984 U.S. Congressional voting	435	16	2	2	61	y
2	LED display	500	7	2	10	10	n
3	LED display + 17 irrelevant attributes	500	24	2	10	10	n
4	Tic-Tac-Toe endgame	958	9	3	2	65	n
5	Standardized audiology	226	69	2	24	48	y
6	Lung cancer data	32	56	3	3	41	y
7	E. coli promoter gene sequences (DNA)	106	57	4	2	50	n
8	Primate splice-junction gene sequences (DNA)	1200	60	4	3	50	n
9	Zoo	101	16	2	7	41	n
10	Hayes-Roth & Hayes-Roth (1977)	160	4	4	3	41	n

Fig. 1. A statistic overview of the examples.

single example would contain many values that practically imply that this attribute or this symptom is missing, and is not needed for the classification of the patient. From the examples presented by Porter et al. [7] it seems that only about a dozen attributes are actually used by a human expert to diagnose each patient. Databases #7 and #8 are taken from molecular biology. Each example in them describes a sequence of nucleotides in the DNA. The learning algorithm relates to each nucleotide as an attribute. Some of the nucleotides sequences (i.e., the examples) are responsible for a certain biochemistry activity (e.g., the production of protein). The task of the algorithm is to identify whether a specific sequence (i.e., a certain example) initiates this activity, and thus belongs to the desired concept, or not. It is known that some of the nucleotides in each sequence, (e.g., some of the attributes in the example) are the ones that evidence whether the sequence initiates the activity or not (e.g., whether the example belongs to the target concept or not), but in different examples different nucleotides are the ones that determine the behavior of the sequence, in other words, in different examples different attributes evidence on membership of the desired concept. Therefore one needs a large set of attributes though for each example only a subset of them actually determines the type of the example. We may conclude that in some domains a large number of attributes is needed, though in each examples some of the attributes might be irrelevant

Some of the domains contain missing values. Missing attribute values degrade the performance of a learning algorithm; yet some algorithms (e.g., ID3) are more

susceptible to missing attributes than others, or make certain assumptions on the distribution or prevalence of the missing values. Quinlan, calls the percent of the missing values ‘the ignorance level’. He writes: “in practice, an ignorance level of even 10% is unlikely” ([8], p. 99). Naturally, Quinlan’s decision trees perform poorly on databases that contain many missing values. EMBLEMs are generally less susceptible to missing values than decision trees, thus we can expect that the performance of an EMBLEM would degrade less than the performance of a decision tree based algorithm when the database contains missing attributes.

5.2 A Comparison with Other Similarity Measures

A similarity measure as described above is only a tool that can be used by a learning algorithm, and composes a central component in any EMBLEM. In this section the similarity measure is relatively isolated from other possible components of a learning algorithm in order to evaluate it on its own.

A Nearest Neighbour Classifier. The most simple examination of the Context-Similarity in a learning task would probably be to incorporate it in a nearest-neighbour classifier, and compare the performance of this classifier with those of one that uses another measure. Such a comparison was made with four other measures. All the comparisons were performed in the same manner: Each classifier received 70% of the examples as a training set, and was tested on the remaining 30% of the examples.^{1 2} All the classifiers were tested on the same training set and test set. For each classifier and each domain fifty runs were executed. Figure 2 depicts the results of this experiment.

If we examine the figure we notice that in eight out of the ten domains the Context-Similarity performs better than the City-Block one. In some of these domains the difference is greater than in others. Whether the difference is meaningful or not is probably task dependent. As the City-Block measure is cheaper to compute, in each implementation it should be considered whether it is worth using the more expensive Context-Similarity measure or not.

A comparison between the Context-Similarity and the Contrast-Similarity shows that in almost all databases the former performs better. Shanon describes the Contrast-Similarity as “the one that defines the state of the art in the field [of cognitive similarity measures]” ([10], p. 308). If one accepts Shanon’s statement then the above finding may seem surprising: How does it happen that a computational model performs better than a cognitive one? I may try to suggest few possible answers to this question, (the different answers do not necessarily exclude each other):

¹ Exceptions are the Audiology database and the Hayes-Roth and Hayes-Roth database. In these two databases the examples are divided into standard training set and test set, therefore, a single run was performed with each of them

² If a test item was most similar to n training set examples, i.e., its similarity to all of the n examples was equal, m out of them from the correct concept, and $n - m$ from wrong ones, then its classification level was taken to be m/n

Average classification rates

Domain #	Domain	Average classification rate of Context-Similarity	Average classification rate of VDM	Average classification rate of WFM	Average classification rate of Contrast-Similarity	Average classification rate of City-Block
1	1984 U.S. Congressional voting	.92	.92	.94	.95	.95
2	LED display	.61	.60	.60	.60	.60
3	LED display + 17 irrelevant attributes	.48	.50	.57	.68	.65
4	Tic-Tac-Toe endgame	.68	.49	.56	.60	.61
5	Standardized audiology	.76	.84	.88	.80	.85
6	Lung cancer data	.32	.37	.36	.42	.45
7	E. coli promoter gene sequences (DNA)	.79	.80	.80	.86	.84
8	Primate splice-junction gene sequences (DNA)	.70	.71	.72	.88	.80
9	Zoo	.95	.96	.95	.95	.97
10	Hayes-Roth & Hayes-Roth (1977)	.87	.90	.93	.94	.99

Number of domains in which performs better than CS	2	0	1	3
Number of domains in which performs less well than CS	8	9	8	5

Fig. 2. The classification rates of five different nearest neighbour classifiers, on a set of ten domains. Each classifier is based on a different similarity measure. For each classifier a comparison with the Context-Similarity based algorithm is also presented: in how many domains this classifier performs better (less well) than the Context-Similarity based one

- The Contrast-Similarity models human similarity scaling well, but humans do not represent objects by vectors of a fixed size. Actually, when Tversky proposed his model one of his main arguments was that people represent objects by *sets of features*; different objects are represented by different sets (and not by a fixed set, as is usually done in machine learning).
- The contrast model describes objects by sets of qualitative features. These features might be either identical or different, but there are no graded degrees of similarity between features.

Stanfill and Waltz, in their MBRtalk system [11], proposed two new similarity measures: the *Weighted Feature Metric (WFM)*, and the *Value Differences Metric (VDM)*.

As can be seen in Fig. 2 the Context-Similarity performs better than WFM in eight out of the ten domains, and better than VDM in five of them. The difference between WFM and VDM is that the latter evaluates separately the similarity between every pair of values (thus allowing v to have different distance from u than from w), while the former assume that the similarity of a value to all

other values is equal (i.e., the distance of v from u is equal to its distance from w). We may, therefore, conclude that a more 'fine' similarity evaluation, one that assigns a unique similarity value to each pair of values, improves performance.

The three main differences between VDM and the Context-Similarity are that the former gives weights to attributes, a property that does not exist in the latter; on the other hand, the latter evaluates the similarity between any pair of values *with respect to each concept separately*, while in the former the similarity between any pair of values is not dependent upon the concept to which the examples belong. Thirdly, in VDM the contribution of two equal values to the similarity between the objects they describe is equal for all values, while in the Context-Similarity each pair of equal values has a unique contribution to the overall similarity between their objects.

It is difficult to characterize the domains in which each measure is superior, or to generalize from these results when should we use one similarity measure or another. Yet, if we try, tentatively, to analyze the results, we note that on the two molecular biology domains (#7 and #8) and on the 'LED display + 17 irrelevant attributes' database VDM performs best. These domains contain *many* irrelevant attributes. It, therefore, seems that in such domains VDM's feature weighting approach is more successful than the Context-Similarity's concept dependent method (i.e., a method that considers the specific concept when evaluating similarity).

Now, turn to the following items: the 'zoo' database, which composes 'classic' natural concepts, and the 'Hayes-Roth and Hayes-Roth' example, that is intended to be an artificial natural concept. On these databases the Context-Similarity performs best. These results hints that the Context-Similarity suites better than other measures for natural concepts. This conclusion is not definite as for example in the 'LED display' domain, which is essentially similar to the 'Hayes-Roth and Hayes-Roth' example, the Context-Similarity does not perform better than the other measures.

Figures 3 and 4 compares between a Context-Similarity based nearest neighbour classifier and two others (VDM and City-Block) on two different domains ('1984 U.S. Congressional voting' and 'E. Coli promoter gene sequences (DNA)'). the figures present the classification rates of each classifier as a function of the size of the training set. The two main findings that are presented in these figures are: (a) The accuracy of both classifiers improve gradually as the size of the training set grows. (b) Generally, if one classifier performs better than another, it does so for all sizes of training sets and test sets; in other words, in most domains it is not the case that while for certain sizes of training sets one classifier performs better, for other sizes another classifier is more accurate.

Overall Similarity. The experiment reported above involved a nearest neighbour classifier. The performance of this classifier is affected by the similarity of a test item E to a single example in each concept—the example that is most similar to E ; the similarity of E to other members of each concept does not affect the prediction of the classifier. The experiment demonstrates a common

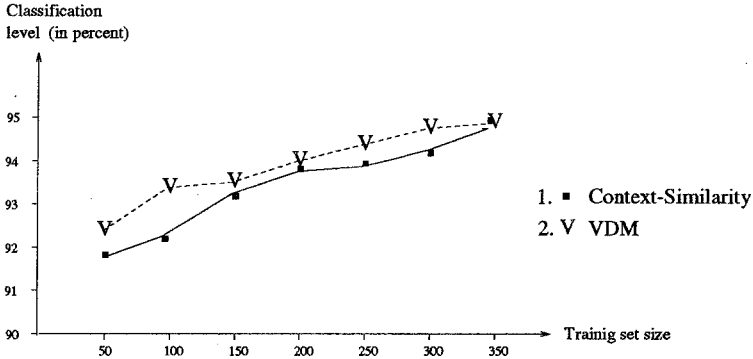


Fig. 3. The classification rates of a Context-Similarity based nearest neighbour classifier versus a VDM based nearest neighbour classifier on the domain '1984 U.S. Congressional voting'. The x axis denotes the size of the training set, the y axis denotes the classification rate of each classifier.

usage of a similarity measure in learning tasks, yet it examines only one aspect of the similarity measure.

In order to examine another aspects of a similarity measure I propose another experiment: A common assumption in clustering and concept learning is that members of a same concept are relatively similar to each other, while members of contrasting concepts are relatively dissimilar. The extent that this assumption is valid varies across domains; some domains satisfy it well, as each of their concepts is centered in a definable region that is well separated from regions occupied by other concepts; other domains satisfy this assumption more loosely either as their concepts are composed of more than a single cohesive and distinct cluster, or as the separability of the concepts is less sharp. Yet, to some extent, this assumption holds, especially if, using some clustering method, we divide

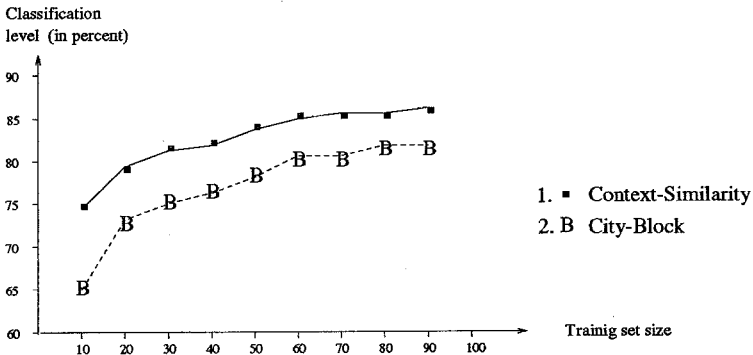


Fig. 4. The classification rates of a Context-Similarity based nearest neighbour classifier versus a City-Block based nearest neighbour classifier on the domain 'E. Coli promoter gene sequences (DNA)'. The x axis denotes the size of the training set, the y axis denotes the classification rate of each classifier.

each concept to a set of sub-concepts in a way that aims to create cohesive and distinct categories.

If we accept this ‘intra concept similarity versus inter concept dissimilarity’ assumption, then we would prefer a similarity measure that reflects this property best. Therefore, this could be another criterion for evaluating a similarity measure. To quantify the extent that a similarity measure expresses this property of a set of concepts, I propose a measure that divides the average similarity between pairs of examples that belong to a same concept, by the average similarity between pairs of examples that belong to contrasting concepts. A larger value of this measure is preferred, as it evidences that the average similarity *within* a concept is large, while the average similarity *between* concepts is small. The *wb* measure could therefore be defined as:

$$wb ::= \frac{\frac{1}{n} \sum_C \sum_{E_i, E_j \in C} sim(E_i, E_j)}{\frac{1}{m} \sum_{C_i} \sum_{C_j \neq C_i} \sum_{E_k \in C_i, E_l \in C_j} sim(E_k, E_l)}$$

where C , C_i , etc. denote concepts, E_i etc. denote examples, $sim(E_i, E_j)$ etc. denote the similarity between E_i and E_j as evaluated by the similarity measure that is examined, n and m denote the number of elements that are summed in the two summations.


This measure does not take into account the variance of the similarity scores of the examined similarity measure. One may argue that if two similarity measures score the same grade in *wb*, yet one of them produces more homogeneous evaluations than the other, then the first is preferred. Therefore to account for the variance of the scores each similarity measure produces, the numerator and denominator of the above *wb* measure are divided by the standard deviation of the similarity scores of the within/between pairs respectively.

The five similarity measures that are used in the experiments were examined using the *wb* measure on the ten domains that were presented before. Figure 5 presents the results of this experiment. As expected, the *wb* scores of most measures in most domains is larger than one, which evidence that, as expected, the evaluated intra concept similarity is larger than the inter concept similarity. This result is in accordance with the ‘intra concept similarity versus inter concept dissimilarity’ assumption, and thus strengthen our confidence in the assumption and the proposed experiment.

If we calculate the average *wb* scores for each similarity measure on the ten domains, we obtain the following results: Contrast Similarity—1.18, City Block—1.32, WFM—1.46, VDM—1.63, Context Similarity—1.67. This ordering of the measures resembles the ordering from the previous experiment. Thus, again, we see that the Context-Similarity performs best, though only slightly better than VDM.

The results reported above were obtained on the original domains. As was mentioned earlier, some of the domains that are used are not composed of concepts that comprise a single distinct cluster. In a similar experiment the original domains were first subject to a clustering program (a variant of the k -means method). This version of the experiment was composed of two phases: In the

The ratio: within concept similarity to between concept similarity



1	1984 U.S. Congressional voting	1.73	1.82	2.20	1.79	2.76
2	LED display	1.99	1.84	1.84	1.61	2.20
3	LED display + 17 irrelevant attributes	1.21	1.15	1.41	1.16	1.80
4	Tic-Tac-Toe endgame	1.02	1.06	1.12	1.00	1.05
5	Standardized audiology	1.12	1.46	1.51	0.69	1.28
6	Lung cancer data	1.24	0.93	0.95	0.97	1.29
7	E. coli promoter gene sequences (DNA)	0.86	1.40	1.33	0.93	0.70
8	Primate splice-junction gene sequences (DNA)	0.96	1.26	1.25	0.96	0.83
9	Zoo	1.98	2.53	3.07	1.66	3.88
10	Hayes-Roth & Hayes-Roth (1977)	1.17	1.21	1.62	1.10	0.93

Each item in this table presents the following measure for some similarity measure in a certain domain:

Average similarity of pairs of items that belong to the same concept
Standard deviation of the similarity scores of items the belong to the same concept

Average similarity of pairs of items that belong to contrasting concepts
Standard deviation of the similarity scores of items the belong to contrasting concepts

Fig. 5.

first phase the clustering program was executed on each domain; the program divided each concept into a set of homogeneous and distinct sub-concepts each having a unique label. The output of the clustering program served as the input to a second phase that was identical to the original experiment (as described in the preceding section). The results of this experiment were even more in favor of the Context-Similarity: Its score on the *wb* measure was improved from 1.67 to 2.12, and the gap between it and VDM grew from 0.04 to 0.52.

As aforesaid, the difference between this experiment and the previous one is that while in the former the results are dependent on the similarity of a probe item to a single example in each concept, in the latter all the possible pairs enter into the calculated measure. Though the two experiments differ in some aspects they both evaluate the same similarity measures using the same domains. We may raise the question what is the correspondence between the results of the two experiments? To answer this question I calculated the Pearson correlation between the average accuracy of the five classifiers in each domain, and the average *wb* ratios of the corresponding similarity measures on these domains. The Pearson correlation equals 0.46. The fact that the correlation is positive confirms the intuition that there is, or there should be, a correspondence between the two aspects that are examined. The finding that the correlation is not very large may be interpreted as an evidence that the two experiments do not examine the same property.

6 Conclusions

This paper presented a context similarity measure, and compared it with other measures. Not surprisingly, in some cases the Context-Similarity measure performs better than other measures, while in other examples it was less accurate. As a conclusion of this paper I would like to suggest that future research should try to either combine the different measures, or characterize more thoroughly which similarity measure is best suited for which class of concepts.

Acknowledgements

I am grateful to Eliezer L. Lozinskii for helpful comments on draft of this paper. I also want to thank Murphy, P. M., & Aha, D. W. who are in charge of the *UCI Repository of machine learning databases* [Machine-readable data repository]; Irvine, CA: University of California, Department of Information and Computer Science. and to professor Jergen at Baylor College of Medicine, who created the audiology database.

References

1. D. W. Aha and D. Kibler. Noise-tolerant instance-based learning algorithm. In *Proc. of the 11'th Int. Joint Conf. on AI*, pages 794–799. Morgan Kaufmann, 1989.
2. D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.
3. J. R. Anderson and M. Matessa. Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning*, 9:275–308, 1992.
4. D. L. Hintzman. “schema abstraction” in a multiple trace memory model. *Psychological Review*, 93:411–428, 1986.
5. D. L. Medin and M. M. Schaffer. Context theory of classification learning. *Psychological Review*, 85:207–238, 1978.
6. R. S. Michalski, R. E. Stepp, and E. Diday. A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts. In L. N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition (Volume 1)*, pages 33–56. North-Holland, New York, 1981.
7. B. W. Porter, R. Bareiss, and R. C. Holte. Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence*, 45:229–263, 1990.
8. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
9. S. Salzberg. A nearest hyperrectangle learning method. *Machine Learning*, 6:251–276, 1991.
10. B. Shanon. On the similarity of features. *New ideas in psychology*, 8:307–321, 1988.
11. C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, 29:1213–1228, 1986.
12. A. Tversky. Features of similarity. *Psychological review*, 84:327–352, 1977.