A Similarity Measure for Clustering and its Applications

Guadalupe J. Torres, Ram B. Basnet, Andrew H. Sung, Srinivas Mukkamala, and Bernardete M. Ribeiro

Abstract—This paper introduces a measure of similarity between two clusterings of the same dataset produced by two different algorithms, or even the same algorithm (K-means, for instance, with different initializations usually produce different results in clustering the same dataset). We then apply the measure to calculate the similarity between pairs of clusterings, with special interest directed at comparing the similarity between various machine clusterings and human clustering of datasets. The similarity measure thus can be used to identify the best (in terms of most similar to human) clustering algorithm for a specific problem at hand. Experimental results pertaining to the text categorization problem of a Portuguese corpus (wherein a translation-into-English approach is used) are presented, as well as results on the well-known benchmark IRIS dataset. The significance and other potential applications of the proposed measure are discussed.

Keywords—Clustering Algorithms, Clustering Applications, Similarity Measures, Text Clustering.

I. INTRODUCTION AND MOTIVATION

Our study of similarity of clustering was initially motivated by a research on automated text categorization of foreign language texts, as explained below.

As the amount of digital documents has been increasing dramatically over the years as the Internet grows, information management, search, and retrieval, etc., have become practically important problems.

Developing methods to organize large amounts of unstructured text documents into a smaller number of meaningful clusters would be very helpful as document clustering is vital to such tasks as indexing, filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of web resources and, in general, any application requiring document organization [1], [2]. Document clustering is also useful for topics such as Gene Ontology [3] in biomedicine where hierarchical catalogues are needed.

To deal with the large amounts of data, machine learning approaches have been applied to perform Automated Text Clustering (ATC). Given an unlabeled dataset, this ATC system builds clusters of documents that are hopefully similar to clustering (classification, categorization, or labeling) performed by human experts.

To identify a suitable tool and algorithm for clustering that produces the best clustering solutions, it becomes necessary to have a method for comparing the results of different clustering algorithms. Though considerable work has been done in designing clustering algorithms, not much research has been done on formulating a measure for the similarity of two different clustering algorithms.

Thus, the main goal of this paper is to: First, propose an algorithm for performing similarity analysis among different clustering algorithms; second, apply the algorithm to calculate similarity of various pairs of clustering methods applied to a Portuguese corpus and the Iris dataset; finally, to cross-validate the results of similarity analysis with the Euclidean (centroids) distances and Pearson correlation coefficient, using the same datasets. Possible applications are discussed.

II. CLUSTERING METHODS

A cluster is a collection of objects which are ‘similar’ between them and are ‘dissimilar’ to the objects belonging to other clusters [4]; and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set.

There are several categories of clustering algorithms. In this paper we will be focusing on algorithms that are exclusive in that the clusters may not overlap.

Some of the algorithms are hierarchical and probabilistic. A hierarchical algorithm clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations, it reaches the final clusters wanted. The final category of probabilistic algorithms is focused around model matching using probabilities as opposed to distances to decide clusters. EM or Expectation Maximization is an example of this type of clustering algorithm.

In [5], Pen et al. utilized cluster analysis composed of 2 methods. In Method I, a majority voting committee with 3 results generates the final analysis result. The performance measure of the classification is decided by majority vote of the committee. If more than 2 of the committee members give the same classification result, then the clustering analysis for that observation is successful; otherwise, the analysis fails.
Kallot et al. [6] did clustering and after letting the algorithm create its own clusters, added a step. After the clustering was completed each member of a class was assigned the value of the cluster’s majority population. The authors noted that the approach loses detail, but allowed them to evaluate each clustering algorithm against the “correct” clusters.

III. THE SIMILARITY MEASURE ALGORITHM

To measure the ‘similarity’ of two sets of clusters, we define a simple formula here: Let C = \{C_1, C_2, …, C_m\} and D = \{D_1, D_2, …, D_n\} be the results of two clustering algorithms on the same data set. Assume C and D are “hard” or exclusive clusters.

Define a simple formula here: Let C = \{C_1, C_2, …, C_m\} and D = \{D_1, D_2, …, D_n\} be the results of two clustering algorithms on the same data set. Assume C and D are “hard” or exclusive clusters.

Thus, the similarity Sim(C, D), according to the similarity matrix above, is \((\sum_{i,j} S_{ij} / \max(m, n))^2\), this would have the effect of giving a lower value of similarity but without changing its range of (0, 1). This similarity measure is a reasonable one to use because, if we define the dissimilarity or “distance” between two clusterings C and D as \(U(C,D) = 1 – \text{Sim}(C,D)\), then it can be proved that \(U(C,D)\) is a good distance measure for it satisfies all desirable properties (non-negativity, identity, symmetry, triangle inequality) of a distance metric.

![Fig. 1 Methodology for calculating similarity measure of clustering a Portuguese corpus. The details of the “translation based text categorization” technique for foreign-language texts are found in [8] and briefly described below.](image)

Fig. 1 illustrates the steps carried out for similarity measure of clustering a Portuguese corpus. The details of the “translation based text categorization” technique for foreign-language texts are found in [8] and briefly described below. (The Iris dataset that is used in our second set of experiments does not require any preprocessing.)

![Fig. 1 Methodology for calculating similarity measure of clustering the Portuguese dataset](image)

A. THE DATASETS

The Portuguese CETEMPublico corpus consisting of 1.5 million extracts, more than 225 million tokens, is excerpts of Portuguese newspaper Público [9]. There are total of 9 different categories which are shown in Table III. The “nd” category which is short for “not defined” was excluded from our experiments. 1000 randomly chosen documents with at least 75 tokens were extracted for each category and then translated into English using the Google translation service [10].

Iris dataset, one of the most popular datasets in pattern recognition literature, was used as benchmark dataset. The dataset can be downloaded from Machine Learning Repository at University of California, Irvine [11]. The dataset is summarized in Table IV.
B. Preprocessing Portuguese Dataset

Tokenization was carried out by using suitable delimiters such as white-space and punctuation marks. Stop words or functional words such as article, prepositions, etc. that are not useful in the text categorization process were removed during preprocessing. Stemming was used to extract the root form of each word in the document. Since stem word as features performs better than single words and noun-phrase [12], we applied the popular and publicly available Porter Stemmer algorithm to stem translated English words [13].

Though there are various term weighting schemes such as BINARY, TF, LOGTF, LOGTFIDF, IDF, TF-CHI, TF-RF [14], [15], we used the traditional but popular weighting scheme, TF.IDF which is one of the best performance wise.

C. Clustering Algorithms

In experimenting with our clustering similarity algorithm the following clustering algorithms were studied:

A. Repeated Bisection
B. Direct
C. Agglomerative
D. Graph
E. K-means
F. K-medoids
G. EM

For the first four algorithms (A - D), gCLUTO [16], a cross-platform graphical application for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters, was used. gCLUTO is built on top of the CLUTO clustering library.

For K-means (E) and K-medians (F) the Matlab Fuzzy Clustering and Data Analysis Toolbox [17] was utilized.

Finally, for Expectation Maximization (G) the WEKA (Waikato Environment for Knowledge Analysis) [18] tool was used.

D. Clustering Similarity Analysis

After applying the clustering algorithms on Portuguese and Iris datasets, clustering similarities were calculated using the proposed algorithm. The results were then verified by calculating centroid Euclidean distance and Pearson correlation.

Euclidean distance:

\[ d = \sqrt{\sum (X - Y)^2} \]  

Pearson correlation coefficient:

\[ r = \frac{\sum x \sum y}{\sqrt{\left( \sum x^2 \right) \left( \sum y^2 \right)}} \]

V. EXPERIMENTAL RESULTS

Pair-wise similarity matrix for all the clustering algorithms mentioned in section IV,C and with the human-labeled actual categories (H) was generated using the Similarity Algorithm we’ve proposed and cross verified with results from Euclidean distance and Pearson correlation.

Due to space limitation only the final similarity matrix between Repeated Bisection and the rest of the algorithms including human-labeled actual categories are shown and is given below. The significant values in the result tables have been emphasized: value closest to 1 in similarity matrix, smallest value for Euclidian distance i.e. smallest distance between cluster centroids, and closest value to +1 for Pearson correlation i.e. best positive relationship between centroids.

A summary of the similarity among A, B, C, D, and H on the Portuguese Dataset is as shown in Table V. Algorithms E, F, and G were not applied to Portuguese dataset due to their implementation limitation as they ran out of memory on a machine with 4 GB RAM. Repeated Bisection and Agglomerative gave 78% (highest) similar clusters. Repeated Bisection was also most similar to Human-labeled actual categories on Iris dataset is shown in Table VI. Repeated Bisection and Direct algorithms resulted 100% similar clusters while they both gave clusters 95% similar to actual categories. As expected K-means and K-medoids resulted 90% similar clusters, but resulted the clusters that are least similar to human-labeled actual categories.
A. Repeated Bisection (A) vs. Human Categorization (H)

1) Results on Portuguese Dataset

$A_0 - A_7$ are clusters given by Repeated Bisection Algorithm and $H_0 - H_7$ are human-labeled actual categories. The Sim(A, H) is 0.6634.

<table>
<thead>
<tr>
<th>Cl.</th>
<th>H_0</th>
<th>H_1</th>
<th>H_2</th>
<th>H_3</th>
<th>H_4</th>
<th>H_5</th>
<th>H_6</th>
<th>H_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_0</td>
<td>0.0358</td>
<td>0.5544</td>
<td>0.0310</td>
<td>0.0104</td>
<td>0.0229</td>
<td>0.0192</td>
<td>0.0514</td>
<td>0.0305</td>
</tr>
<tr>
<td>A_1</td>
<td>0.0100</td>
<td>0.0110</td>
<td>0.0991</td>
<td>0.0025</td>
<td>0.0030</td>
<td>0.0136</td>
<td>0.5737</td>
<td>0.0100</td>
</tr>
<tr>
<td>A_2</td>
<td>0.0987</td>
<td>0.0088</td>
<td>0.1123</td>
<td>0.0056</td>
<td>0.0139</td>
<td>0.0315</td>
<td>0.0197</td>
<td>0.1217</td>
</tr>
<tr>
<td>A_3</td>
<td>0.0105</td>
<td>0.0033</td>
<td>0.0110</td>
<td>0.7075</td>
<td>0.0043</td>
<td>0.0129</td>
<td>0.0361</td>
<td>0.0183</td>
</tr>
<tr>
<td>A_4</td>
<td>0.0379</td>
<td>0.0258</td>
<td>0.0105</td>
<td>0.0048</td>
<td>0.5387</td>
<td>0.0177</td>
<td>0.0013</td>
<td>0.1372</td>
</tr>
<tr>
<td>A_5</td>
<td>0.2781</td>
<td>0.0481</td>
<td>0.0359</td>
<td>0.0058</td>
<td>0.0245</td>
<td>0.0235</td>
<td>0.0043</td>
<td>0.1779</td>
</tr>
<tr>
<td>A_6</td>
<td>0.0472</td>
<td>0.0037</td>
<td>0.3282</td>
<td>0.0062</td>
<td>0.0150</td>
<td>0.0444</td>
<td>0.0100</td>
<td>0.0150</td>
</tr>
<tr>
<td>A_7</td>
<td>0.0639</td>
<td>0.0010</td>
<td>0.0173</td>
<td>0.0045</td>
<td>0.0163</td>
<td>0.0535</td>
<td>0.0116</td>
<td>0.0433</td>
</tr>
</tbody>
</table>

2) Results on Iris Dataset

Repeated Bisection (A) did show a slight deviation from a perfect match to human categorization. Clusters A_1 and H_1 showed a distance of 0.1073 and A_2 and H_1 0.0643. Clusters centroid A_0 showed a perfect match with the human labeled centroid H_0. The similarity values of the Pearson correlation coefficient support this as they range from 0.99992-1. This supports the 95% similarity result obtained from our similarity algorithm.

<table>
<thead>
<tr>
<th>Cl.</th>
<th>B_0</th>
<th>B_1</th>
<th>B_2</th>
<th>B_3</th>
<th>B_4</th>
<th>B_5</th>
<th>B_6</th>
<th>B_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_0</td>
<td>0.4199</td>
<td>0.5278</td>
<td>0.5395</td>
<td>0.3436</td>
<td>0.9593</td>
<td>0.4722</td>
<td>0.4487</td>
<td>0.3171</td>
</tr>
<tr>
<td>C_1</td>
<td>0.4533</td>
<td>0.6422</td>
<td>0.9672</td>
<td>0.3128</td>
<td>0.4502</td>
<td>0.4416</td>
<td>0.3643</td>
<td>0.4783</td>
</tr>
<tr>
<td>C_2</td>
<td>0.4753</td>
<td>0.6583</td>
<td>0.3864</td>
<td>0.3172</td>
<td>0.3585</td>
<td>0.6476</td>
<td>0.4271</td>
<td>0.6333</td>
</tr>
<tr>
<td>C_3</td>
<td>0.4254</td>
<td>0.3940</td>
<td>0.3790</td>
<td>0.3964</td>
<td>0.3372</td>
<td>0.4133</td>
<td>0.3856</td>
<td>0.4409</td>
</tr>
<tr>
<td>C_4</td>
<td>0.4843</td>
<td>0.4882</td>
<td>0.3716</td>
<td>0.3938</td>
<td>0.4738</td>
<td>0.7206</td>
<td>0.9543</td>
<td>0.6168</td>
</tr>
<tr>
<td>C_5</td>
<td>0.5540</td>
<td>0.6611</td>
<td>0.4793</td>
<td>0.4057</td>
<td>0.5438</td>
<td>0.7754</td>
<td>0.6166</td>
<td>0.8795</td>
</tr>
<tr>
<td>C_6</td>
<td>0.5329</td>
<td>0.6055</td>
<td>0.4339</td>
<td>0.3208</td>
<td>0.3691</td>
<td>0.4520</td>
<td>0.4144</td>
<td>0.5766</td>
</tr>
<tr>
<td>C_7</td>
<td>0.9479</td>
<td>0.5524</td>
<td>0.4062</td>
<td>0.3782</td>
<td>0.3326</td>
<td>0.5237</td>
<td>0.4279</td>
<td>0.5685</td>
</tr>
</tbody>
</table>

B. Repeated Bisection (A) vs. Direct (B)

1) Results on Portuguese Dataset

$A_0 - A_7$ are clusters given by Repeated Bisection and $B_0 - B_7$ are clusters given by Direct algorithms. The Sim(A, B) is 0.7813.

2) Results on Iris Dataset

Clusters $A_0 - A_7$ are clusters given by Repeated Bisection and $B_0 - B_7$ are clusters given by Direct clustering algorithm. The similarity between Repeated Bisection and Direct is 1, suggesting 100% similarity between the clusters given by A and B, which infers that Repeated Bisection and Direct algorithms gave clusters with 100% similarity.

<table>
<thead>
<tr>
<th>Cl.</th>
<th>B_0</th>
<th>B_1</th>
<th>B_2</th>
<th>B_3</th>
<th>B_4</th>
<th>B_5</th>
<th>B_6</th>
<th>B_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_0</td>
<td>0.8303</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0005</td>
<td>0.0058</td>
<td>0.0605</td>
<td>0.0074</td>
<td>0.0000</td>
</tr>
<tr>
<td>A_1</td>
<td>0.0032</td>
<td>0.8965</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0144</td>
<td>0.0038</td>
<td>0.0000</td>
</tr>
<tr>
<td>A_2</td>
<td>0.0029</td>
<td>0.0030</td>
<td>0.0081</td>
<td>0.5000</td>
<td>0.2019</td>
<td>0.0187</td>
<td>0.0315</td>
<td>0.0110</td>
</tr>
<tr>
<td>A_3</td>
<td>0.0017</td>
<td>0.0000</td>
<td>0.9402</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0031</td>
<td>0.0062</td>
<td>0.0004</td>
</tr>
<tr>
<td>A_4</td>
<td>0.0012</td>
<td>0.0000</td>
<td>0.0008</td>
<td>0.0009</td>
<td>0.5139</td>
<td>0.0148</td>
<td>0.2877</td>
<td>0.0004</td>
</tr>
<tr>
<td>A_5</td>
<td>0.0045</td>
<td>0.0014</td>
<td>0.0022</td>
<td>0.2043</td>
<td>0.0006</td>
<td>0.5542</td>
<td>0.0469</td>
<td>0.0056</td>
</tr>
<tr>
<td>A_6</td>
<td>0.0151</td>
<td>0.0042</td>
<td>0.0057</td>
<td>0.2234</td>
<td>0.0018</td>
<td>0.0345</td>
<td>0.0969</td>
<td>0.0596</td>
</tr>
<tr>
<td>A_7</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0009</td>
<td>0.0022</td>
<td>0.0059</td>
<td>0.0072</td>
<td>0.0250</td>
<td>0.8179</td>
</tr>
</tbody>
</table>

The comparison of the results of the Repeated Bisection and Direct clustering algorithms showed perfect matches with each other once again supporting our 95% similarity comparison result.
C. Repeated Bisection (A) vs. Agglomerative (C)

1) Results on Portuguese Dataset

A₀ - A₇ are clusters given by the Repeated Bisection and C₀ - C₇ are clusters given by Agglomerative algorithm. The Sim(A, C) is 0.6078.

2) Results on Iris Dataset

Clusters A₀ - A₂ are clusters given by Repeated Bisection and C₀ - C₂ are clusters given by Agglomerative clustering algorithm. The Sim(A, C) is 0.9360. Observe that clusters A₀ and C₀ are 100% similar; however clusters A₁ and C₁, and A₂ and C₂ are 87% and 86% similar respectively, which brought the average similarity down to 93% compared to Repeated Bisection and Direct.

The comparison of the results of the Repeated Bisection and Agglomerative clustering algorithm showed near perfect matches with each other supporting our 94% similarity comparison result.

D. Repeated Bisection (A) vs. Graph (D)

1) Results on Portuguese Dataset

A₀ - A₇ are clusters given by Repeated Bisection algorithm and D₀ - D₇ are clusters given by Graph algorithm. The Sim(A, D) is 0.5963.

2) Results on Iris Dataset

Clusters A₀ - A₂ are clusters given by Repeated Bisection and D₀ - D₂ are clusters given by Graph clustering algorithms. The Sim(A, D) is 0.7425. Notice that Graph algorithm suggested 4 clusters instead of requested 3; which significantly brought the overall similarity to 74% though there are two cluster sets that are as high as 100% similar.

The comparison of the results of the Repeated Bisection and Graph clustering algorithm showed differences with each other supporting our lower 72% similarity comparison result.

E. Repeated Bisection (A) vs. K-means (E)

The comparison of the results of the Repeated Bisection and K-means clustering algorithm showed significant difference with each other once again supporting our 67% similarity.
each supporting the 84% similarity comparison result, as it
EM clustering algorithm showed significant difference with
B, or C.

TABLE XXV
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND E

<table>
<thead>
<tr>
<th>Cl.</th>
<th>E₀</th>
<th>E₁</th>
<th>E₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>0.6700</td>
<td>4.0561</td>
<td>0.2635</td>
</tr>
<tr>
<td>A₁</td>
<td>4.4354</td>
<td>0.6187</td>
<td>4.5921</td>
</tr>
<tr>
<td>A₂</td>
<td>2.8878</td>
<td>0.9544</td>
<td>3.1245</td>
</tr>
</tbody>
</table>

TABLE XXV
PEARSON CORRELATION BETWEEN A AND E

<table>
<thead>
<tr>
<th>Cl.</th>
<th>E₀</th>
<th>E₁</th>
<th>E₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>0.99902</td>
<td>0.6858</td>
<td>0.9997</td>
</tr>
<tr>
<td>A₁</td>
<td>0.7238</td>
<td>0.9964</td>
<td>0.6074</td>
</tr>
<tr>
<td>A₂</td>
<td>0.8499</td>
<td>0.9924</td>
<td>0.7567</td>
</tr>
</tbody>
</table>

F. Repeated Bisection (A) vs. K-medoids (F)
The comparison of the results of the Repeated Bisection and
K-medoids clustering algorithm showed significant difference with
each other once again supporting our 66% similarity comparison result.

TABLE XXVI
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND F

<table>
<thead>
<tr>
<th>Cl.</th>
<th>F₀</th>
<th>F₁</th>
<th>F₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>0.3340</td>
<td>4.0372</td>
<td>0.5154</td>
</tr>
<tr>
<td>A₁</td>
<td>4.6003</td>
<td>0.6400</td>
<td>4.5303</td>
</tr>
<tr>
<td>A₂</td>
<td>3.1419</td>
<td>0.9325</td>
<td>2.9908</td>
</tr>
</tbody>
</table>

TABLE XXVII
PEARSON CORRELATION BETWEEN A AND F

<table>
<thead>
<tr>
<th>Cl.</th>
<th>F₀</th>
<th>F₁</th>
<th>F₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>0.9996</td>
<td>0.6802</td>
<td>0.9956</td>
</tr>
<tr>
<td>A₁</td>
<td>0.6016</td>
<td>0.9964</td>
<td>0.6925</td>
</tr>
<tr>
<td>A₂</td>
<td>0.7520</td>
<td>0.9924</td>
<td>0.8255</td>
</tr>
</tbody>
</table>

G. Repeated Bisection (A) vs. EM (G)
The comparison of the results of the Repeated Bisection and
EM clustering algorithm showed significant difference with
each supporting the 84% similarity comparison result, as it
was not quite as bad as K-means, nor as good as algorithms A,
B, or C.

TABLE XXVIII
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND G

<table>
<thead>
<tr>
<th>Cl.</th>
<th>G₀</th>
<th>G₁</th>
<th>G₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>3.3668</td>
<td>4.9992</td>
<td>0.0000</td>
</tr>
<tr>
<td>A₁</td>
<td>1.3707</td>
<td>0.4098</td>
<td>4.6557</td>
</tr>
<tr>
<td>A₂</td>
<td>0.2328</td>
<td>1.9494</td>
<td>3.1561</td>
</tr>
</tbody>
</table>

TABLE XXIX
PEARSON CORRELATION BETWEEN A AND G

<table>
<thead>
<tr>
<th>Cl.</th>
<th>G₀</th>
<th>G₁</th>
<th>G₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>0.7355</td>
<td>0.6174</td>
<td>1.0000</td>
</tr>
<tr>
<td>A₁</td>
<td>0.9878</td>
<td>0.9999</td>
<td>0.6227</td>
</tr>
<tr>
<td>A₂</td>
<td>0.9986</td>
<td>0.9773</td>
<td>0.7695</td>
</tr>
</tbody>
</table>

VI. Conclusion
The similarity measure that we proposed has experimentally
demonstrated consistently similar results to popular measures of
Euclidian distance (between cluster centroids) and Pearson
correlation. The measure provides the benefit of allowing the
aggregated comparison between differing algorithms to allow users to identify the best available clustering algorithm for
their applications. Our results show that Repeated Bisection
and Direct hierarchical clustering algorithms consistently
produced clusters that are most similar to expert labeled
categories for both smaller data sets with fewer features (Iris)
and large dataset (translated Portuguese-English corpus) with
a much larger number of features. Though there remains much
room for additional research, our preliminary results indicate
that Repeated Bisection and Direct algorithms can be used for
clustering both small and large scale datasets, for example,
foreign language text document clustering.

With the intriguing initial results, our future work will
include expansion and verification of the proposed algorithm
through the use of larger datasets along with various feature
sizes, sample sizes, and expansion with the numbers of
categories to work with.
The techniques can be extended to various real-world
problems such as classification and clustering of malware,
email analysis (finding social graph among the users based on
email contents, for instance) in digital forensics. Since
unsupervised clustering algorithms do not give accuracy; the
proposed algorithm can be applied to find the best clustering
algorithm for many real-life applications where clustering
techniques are applied. The approach should enable users to
experimentally compare various clustering algorithms and
choose the one that best serves the problem.

ACKNOWLEDGMENT
The authors would like to thank Messrs. D. Chen, R. Lopes,
R. Koduru, S. Mungala, K. Bondili, and M. Batta for their
assistance in translating the Portuguese documents.

REFERENCES
W. E. Hammond, “Medical Data Mining: Knowledge Discovery in a
Clinical Data Warehouse,” American Medical Informatics Association
Methods to Gene Ontology Annotation,” Proceedings of the 28th
Annual International ACM SIGIR Conference on Research and
http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/.
for Credit Card Accounts Classification,” LNCS 3516, 2005, pp. 548-
553.
Supervised Learning, and Data Assignment,” Proceedings of the
Seventh International Conference on Knowledge Discovery and Data
http://people.revoledu.com/kardi/tutorial/Similarity/.
Ribeiro, “Translation Based Arabic Text Categorization,” Proceedings
of Second International Conference on Information Systems Technology
and Management, Dubai, March 2008.
/Repositorio/.