Author’s response to reviews

Title: Evaluation of standard and semantically-augmented distance metrics for neurology patients

Authors:
Daniel B Hier (dhier@uic.edu;dbhier@dbhier.com)
Jonathan Kopel (jjkbqb@mst.edu)
Steven U Brint (sbrint@uic.edu)
Donald C Wunsch II (dwunsch@mst.edu)
Gayla R Olbricht (OlbrichtG@mst.edu)
Sima Azizi (sacc5@mst.edu)
Blaine Allen (bh9z8@umsystem.edu)

Version: 2 Date: 18 Jul 2020

Author’s response to reviews:

To: Mohammad Reza Daliri, PhD
BMC Medical Informatics and Decision Making

Re: MIDM-D-20-00148R1
Evaluation of standard and semantically-augmented distance metrics for neurology patients
Daniel B Hier, MD; Jonathan Kopel; Steven U Brint, MD; Donald C Wunsch II, PhD; Gayla R Olbricht, PhD; Sima Azizi; Blaine Allen

Date: July 16, 2020

Dear Dr. Daliri,

We thank the reviewers for their careful review of our paper. They have offered many helpful suggestions for improvements. We have tried to implement as many of the suggestions as possible; while working within the constraints of the time available and working to keep the paper at a manageable length. We have added an enhanced discussion of the classification and clustering algorithms used in this study. We have also enlarged the background section to more fully discuss prior work on semantically augmented patient distance metrics.

I have outlined our specific responses to each of the reviewers' comments. Again, I thank them for their thoughtful review of our work.
Reviewer 1(a):
I am satisfied with the reviewer corrections. They addressed most of the issues I raised. Obviously, the paper mainly ended up showing no improvement in semantic-based distance measures in terms of classification, a negative result, but I think this reads as better science.
Response to Reviewer 1.a: Thank you for your careful review.

Reviewer 2(a):
In the revised version of the paper, authors have made substantial changes that involve the very hypothesis handled in the original version. As proposed by the reviewers, authors have evaluated other well-known metrics and ML algorithms to compare them with the proposed augmented ones (which was the main novelty in the original paper). Unfortunately, results show no gain in using semantic augmentation, and therefore there is no clear conclusion of this. Traditional metrics have been widely used over text and annotated text, so there is no novelty in showing analytics and classification results with these metrics. Semantic augmentation was expected to improve them, but results showed no improvement. Probably the dataset is too small to lead to sound conclusions, or perhaps semantic augmentation doesn't work because of deficiencies in knowledge resources. Anyway, the paper, in its current form, doesn't contribute anything to the field. As pointed out by the authors, a promising path to be explored is the application of modern NLP based on embeddings to learn more appropriate metric spaces for diagnostic.

Response to Reviewer 2(a): We agree that a larger dataset would be valuable. We will add that to the limitation of the study. We believe the path forward for patient similarity studies will involve concept encoding (either SNOMED CT or UMLS) of large patient datasets derived from either published case studies of electronic health records. NLP techniques will hopefully expedite the development of these large datasets from either EHRs or published case studies in the medical literature. We felt this study was important to perform because of the lack of studies in the biomedical literature that have tried to encode with UMLS or SNOMED CT codes the signs and symptoms of patients with known diagnoses—and then measure the distances between these patients. To our knowledge, published studies on patient similarity have depended mainly on codes derived from either discharge summaries or problem lists.

Reviewer 4(a):
The article primarily focuses on the quality improvement of methodologies to accurately classify neurology diagnosis. Article Highlights: * Usage of Semantic augmentations in classification and clustering models and its insignificant performance to improve clusters/classifiers. Comments: Well described methods. The only take away from this paper that I have is that the findings and methodology may help other researchers to be wary of semantic augmentation applications. Although the hypothesized aim was not achieved, authors can describe future proposals on plausible ways-suggestions to enhance the classification accuracy. A couple of points on how this may help or improve clinical activities. Overall the revision addressed the concerns of the previous
round of reviews. The paper mentions SNOMED CT and UMLS, but SNOMED CT's ontological relationships are not used. The NEO ontology, which is referenced, has some specific characteristics that are worth mentioning explicitly in the paper, such as being mono-hierarchic and manually curated with similarity in mind.

Response to Reviewer 4(a): Thank you. We have added comments on how the NEO ontology differs from both UMLS and SNOMED CT. We have also added comments as to implications for improving neurological diagnosis.

Reviewer 4(b): In the discussion, the authors mention other papers that have reported improvements in the clustering of patients and documents. I would suggest including a more detailed description of this work (at least the work on patients, which is more similar to this paper) in the background and related work section, to give the reader better insights into why and when semantically-distances are useful or not.
Response to Reviewer 4(b): We agree and have enlarged the Background section to describe better prior work on using concept similarity to improve the clustering and classification of patients.

Reviewer 4(c): Equation 2 is missing a closing bracket.
Response to Reviewer 4(c): Thank you, corrected.

Reviewer 4(d): The "p" in "python" should be upper case (page 7).
Response to Reviewer 4(d): Corrected. Thank you.

Reviewer 6(a):
The revised version of the manuscript is much clearer than the original version. The authors mentioned that "Distance metrics are frequently used for clustering and classification of patients. Since the performance of machine learning clustering and classification algorithms can be assessed objectively, we hypothesized that the semantic augmentation of distance metrics with inter-concept distances would improve the performance of these algorithms." I would suggest the authors include some literature or background on this before formulating such hypothesis. Justify the learning clustering and classification algorithms chosen in this study. The authors should introduce common machine learning algorithms used for such domain. In the methods section, the machine algorithms need to be referenced, and the authors should at least provide and justify why these algorithms are selected.
Response to Reviewer 6(a):
We agree. We have added wording to the Background section, indicating our reasoning behind the choice of clustering and classification algorithms. In the methods section, we have referenced the algorithms utilized.

Reviewer 6(b): The author's used default parameters for each algorithm. Parameter optimization of the machine learning algorithm is known to improve the model performance, e.g., the number of trees in the RF model. The same goes for the clustering tasks (unsupervised machine learning), to assess the cluster quality. There was no background nor justification of why are these methods both the
agglomerative clustering algorithm (Ward linkage) and the k-means clustering algorithm chosen in this study.

Response to Reviewer 6(b):
We agree hyperparameter optimization is important to ensure the best performance of classifier and clustering algorithms. We have gone back and optimized the hyperparameters for the RF, logistic regression, KNN, k-means clustering, and agglomerative clustering algorithms.

Reviewer 6(c): I suggest the authors improve the introduction section as it is very heavy on distance metrics alone.
Response to Reviewer 6(c): We agree and have added a brief discussion of classification and clustering algorithms as well.

Reviewer 6(d): How did the authors arrive at the four test groups? The justification based on textbooks alone is not sufficient and very vague.
Response to Reviewer 6(d):
We created a convenience dataset of 1028 neurological cases derived by abstracting teaching cases from 26 textbooks of neurology. We wanted to avoid a very large test set with hundreds of possible diagnoses. Since physicians usually have 4-8 diagnoses on a "differential diagnosis" for a given chief complaint—we tried to emulate this by creating four groups of 8 competing diagnoses that acted as a physician's differential diagnosis of a patient's chief complaint. We picked diagnoses with the highest frequency so that there would be enough instances for training the classifiers. In clinical practice, physicians rarely consider more than a few diagnoses as possible explanations for a patient's presenting complaint.

Reviewer 6(e): How did the authors deal with unbalanced classes?
Response to Reviewer 6(e): Although the groups were unbalanced, they were not severely imbalanced. Each group had eight classes, and the largest class was 25.3% in the movement group, 19% in the mental status group, 19% in the weakness group, and 16% in the cranial nerve group. Unreported statistical testing showed all classifiers outperformed a simple majority vote model. By using both the F1 score and the Accuracy score, we considered both true positives and false positives as a further way of compensating for unbalanced groups.

Reviewer 6(f): Stratified random sampling with k-fold is a better approach.
Response to Reviewer 6(f):
We re-ran the validations for the classifiers as both 10-fold cross-validation and stratified random sampling (k=10). F1, precision, and recall were lower with stratified random sampling than with cross-validation; but none of the mean scores were significantly different:

<table>
<thead>
<tr>
<th>Validation</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROSS.</td>
<td>66.7</td>
<td>64.6</td>
<td>66.9</td>
</tr>
<tr>
<td>RANDOM</td>
<td>66.3</td>
<td>62.9</td>
<td>65.0</td>
</tr>
</tbody>
</table>

Correlation coefficients between random sampling and cross folds were 0.99 for precision, 0.98 for recall, and 0.99 for F1. Since the scores were not significantly different, we have elected to retain the original scores.
Reviewer 6(g): Default parameter was used for the machine learning model could that be the reason why semantic augmentation of the distance metrics did not improve the performance of classification and clustering algorithms?
Response to Reviewer 6(g): We thank the reviewer for this suggestion. As above, we rechecked the hyper-parameters as suggested above for the classification and clustering algorithms. We then recalculated all results with the optimized parameters. We did not find significantly different mean accuracies for the classification algorithms or significant different mean quality measures for the clustering algorithms with the optimized algorithm hyper-parameters when compared to the original default parameters. We now report the hyper-parameters we used to obtain the shown results.

Reviewer 6(h):
The authors should justify why the logistic regression classifier and the k-nearest neighbor classifier outperform the naïve Bayes classifier and RF. RF is only using default trees, which is ten.
Response to Reviewer 6(h): In retrospect, choice of RF and NB classifiers were a poor choice since they are not algorithmically designed to take advantage of the patient distance matrix. Also, NB makes the "naïve" assumption of feature independence, which does not hold for patient distances. We have added this comment to the discussion.

Sincerely,

Daniel B. Hier MD
Corresponding Author