UMAP Based Data Validity Evaluation for Artificial Intelligence Systems

Hanseong Son* and Hyemin Lim
Division of Bio-convergence, Joongbu University, 201 Daehak-ro, Geumsan-gun, Chungnam, 312-702, Republic of Korea

*Email: hsson@joongbu.ac.kr

Abstract. Although comparison is one of the most useful data validity evaluation method, it has a few potential drawbacks. One of them is that results from the training or the testing a machine learning model shall be obtained. Another one is that uncertainty due to the machine learning model itself may cause a difficulty in evaluating the data validity. In this paper, a new data validity evaluation method was proposed so that these drawbacks can be made up by the proposed method. The proposed method enables a data validity evaluation to be performed in data collection phase or data pre-processing phase. Uniform Manifold Approximation and Projection (UMAP) offers the methodological background to the proposed method.

1. Introduction
The training data are playing a crucial role in developing an artificial intelligence system. Thus, the validity of data mainly determines the quality of an AI system. This means that the validity of the training data shall be assured through the development process. Howison and his colleagues tried to define validity by explaining the relationship between the collected data and the conceptual structure of a system, recognizing that validity depends on the harmony between certain data collection methods and theoretical constructs [1]. For example, in social network analysis, surveys, and so on, validity was defined by assessing the extent to which a given test/instrumentation measure a theoretical structure [2]. Based on the validity concepts described in these previous researches, Son defined data validity as the extent how effectively the training data for an AI system reflect the conceptual structure of a real-world [3]. With the definition of data validity, it can be said that, if particular training data collected for an AI system work effectively as test data for a different AI system and the two AI systems are designed to reflect the same theoretical construct, the validity of the training data can be assessed as ‘good’.

One of the most effective evaluation methods of the data validity is comparison. For example, the validity of the data used for expert systems are evaluated by comparing the results from an expert system with those from real experts’ knowledge. Statistical methods such as Bayesian models and face testing can be used for the comparison. However, there are a few potential drawbacks in the comparison based data validity evaluation. One of them is that results from the training or the testing a machine learning model shall be obtained. Another one, which is tightly coupled with the previous drawback, is that uncertainty due to the machine learning model itself may cause a difficulty in evaluating the data validity.

This work proposed a new approach for data validity evaluation, supplementing comparison based methods especially in view that the above mentioned drawbacks can be made up by the proposed
method. The proposed method enable a data validity evaluation to be performed in data collection phase or data pre-processing phase. Section 2 introduces Uniform Manifold Approximation and Projection (UMAP) which offers the methodological background to the proposed method. Section 3 introduces an experiment and its results. Based on the experimental results, the data validity evaluation method is proposed in this work. The proposed validity evaluation method is also described in Section 3. This work is concluded with the suggestion of a few future works in Section 4.

2. Uniform Manifold Approximation and Projection (UMAP)

UMAP is a dimension reduction method using an original and powerful learning technique [4]. Dimension reduction represents high dimensional data in a lower dimension with preserving relevant structure. Dimension reduction is an important machine learning technique because it is very useful for visualization and pre-processing steps in the data science process. Thus, powerful dimension reduction algorithms are desired to deal with massive data and cope with the diversity of data. UMAP is one of the dimension reduction algorithms that can help them be applied in a widening range of areas and on increasing sizes of data sets.

UMAP has a solid mathematical foundation and a practical scalability so that it may be well applied to real-world data. Mathematical substructures of UMAP are related to the work of Belkin and Niyogi on Laplacian eigenmaps [5, 6]. UMAP combined Riemannian geometry with David Spivak’s approach to geometric realization of fuzzy simplicial sets [7] in order to address the issue of uniform data distributions on manifolds. One of the very modern dimension reduction method for visualization using the most recent ideas is t-SNE [8, 9]. For visualization quality, UMAP is competing and contending with t-SNE. However, UMAP has a run time performance superior to t-SNE because it has more of the global structure. In addition, by virtue of its topological foundations, UMAP is scalable to significantly larger sizes of data sets than are feasible for t-SNE. Finally, UMAP is viably considered as a general purpose dimension reduction technique for machine learning because it has no computational restrictions on embedding dimension.

UMAP algorithm is very effective to visualization as well as dimension reduction. It is UMAP is based on the important assumption that the data has manifold structure in them. Therefore, UMAP has a tendency to find manifold structure within the dataset having noises as if the human mind tends to find structured constellations among the stars. UMAP visualization is tightly coupled with this tendency and thus it can be said that the visualization is based on the distance between observations. Since UMAP is derived from the axiom that local distance is more important than long range distances, it mainly concerns representing local structure accurately. As the size of sampled data increases, the amount of structure evident from noise becomes less and UMAP becomes more robust.

3. Data Validity Evaluation

3.1. Experiment Design

This work performed an experiment to evaluate how the difference of data distribution affects the accuracy of machine learning models. Firstly, data sets were created for this experiment. The data sets contain 407 documents each of which presents a symptom. All the symptoms are fallen into the four internal medicine categories and the categories are neurology (encoded as ‘0’), nephrology (encoded as ‘1’), pulmonology (encoded as ‘2’), and gastroenterology (encoded as ‘3’). Every document contains a corpus which consists of words that non-expert persons use to express their symptom. Neurology category has 90 documents. Nephrology category has 104 documents. Pulmonology category has 92 documents. Finally, gastroenterology category has 121 documents. It can be roughly said that the data sets are evenly distributed.

Every corpus in the documents is pre-processed to make it appropriate to machine learning algorithms. Korean morphological analysis is then performed by using Konlpy.twitter package. Through the morphological analysis, the documents are tokenized and the tokenized words are vectorised by using a few vectorizers like CountVectorizer and the Term Frequency–Inverse
Document Frequency (TF-IDF) vectorizer. This work used the TF-IDF vectorizer for the experiment. CountVectorizer is used for a comparative analysis in this work. CountVectorizer tokenizes a collection of text documents and builds a vocabulary of known words. Using the vocabulary, it also encodes new documents. To use CountVectorizer, an instance of the CountVectorizer class shall be created firstly. Secondly the ‘fit’ function is called to learn a vocabulary from one or more documents. Finally the ‘transform’ function on one or more documents as needed is called to encode each as a vector. The vector encoded by CountVectorizer represents a length of the entire vocabulary and the number of times in which each word appeared in the document [10]. TF-IDF is also a calculation method for word frequencies. Term Frequency represents how often a particular word appears within a document. Inverse Document Frequency inversely scores how often a certain word appears across documents in a document based manner. TF-IDF is a production of Term Frequency and Inverse Document Frequency. This means that TF-IDF is a word frequency score focusing on the words that are frequent in a document but not across documents, i.e., are more interesting [10].

The encoded vectors by the TF-IDF vectorizer are used with machine learning algorithms for this experiment. Logistic regression, naïve Bayes, random forest and an ensemble algorithm are the four machine learning algorithms introduced for the experiment. Logistic regression is an algorithm that extend linear regression algorithm so as to be used for classification tasks. Linear regression algorithm seeks the correlation of each feature with the output to predict future or absent values. Thus the output variable of linear regression is continuous (e.g., a list of temperature values). On the contrary, that of logistic regression is binary or discrete (e.g., only hot or cold) [11]. This is because logistic regression imposes soft-max operation on the results from linear regression. Naive Bayes is a classification method based on the Bayesian theorem. According to the Bayesian theorem, the posterior knowledge is gained by updating the prior knowledge of an event with the independent probability of each feature that can affect the event [11]. Naïve Bayes classifies data by using the updated posterior knowledge. Random forest is a bagging algorithm to improve the accuracy of the decision tree algorithm. Random forest combines several decision trees and each tree produces its own label. The 'majority vote' method is used to determine which label to return. Using the 'majority vote' method means that the final output of random forest algorithm is the label with the most vote for the classification task. It also means that the average prediction of all the decision trees is the final output of random forest for the regression task. Decision tree model finds patterns from data-feature values and constructs a tree with branches at decision nodes based on the patterns [11]. Decision tree is used for both classification and regression and thus random forest is. Ensemble methods integrate various weak classification models or regression models into a strong classification model or regression model to obtain better accuracy than any of the constituent models alone [12]. The integration methods are bagging and boosting. Bagging consists of sampling input data many times and aggregating outputs from constituent models. Boosting changes weak models into strong ones by using weights. In this experiment, an ensemble method using logistic regression, naïve Bayes, and random forest algorithms is used.

Training is performed by the originally created data sets that are labelled with the four categories. Then training is repeatedly performed by the data sets where a data set is reduced to 1/2, 1/4, and 1/8 for each label. When the neurology labelled data set is lessened to 1/2, the data set is named as neurology 1/2. Every modified data set is named in the same manner. The different test data sets are used to test each trained machine learning model and the corresponding accuracy results are produced.

3.2. Experiment Results and Data Validity Evaluation Method

Table 1 shows the accuracies of the machine learning models trained with the created and modified data sets. While the accuracies of random forest model are the highest for most of data sets, those of naïve Bayes model are the lowest for all the data sets. It should be noted that the accuracy decreases as the size of a data set corresponding to a label lessens. Basically, this is related to the distribution of the data set sizes. As the size of a data set corresponding to a label decreases, the distribution of the data set sizes changes. If the sizes of the original data sets are uniformly distributed, the uniformity decreases due to the size reduction. Thus, the experiment results in Table 1 show that the accuracy is
getting lower as the uniformity decreases. In other words, the uniformity of the data set size distribution positively affects the accuracy of the machine learning models trained by the data sets. It can also be said that a data set distribution can be called ‘uniform’ when the distribution does not decrease the accuracy of a machine learning model where the corresponding data sets are applied for training, according to the experiment results. Hence, in order to ensure a sufficient accuracy of a machine learning model, it is necessary to collect data sets that are ‘uniformly’ distributed for each label. This raises an issue how much uniformly the data sets should be distributed to acquire an enough accuracy of machine learning models. This issue is related to a measure for the uniformity of the data set distribution.

Table 1. Train data sets and their corresponding machine learning model accuracies.

|                      | Logistic Regression | Naive Bayes | Random Forest | Ensemble |
|----------------------|---------------------|------------|---------------|----------|
| Original Data Set    | 0.83                | 0.76       | 0.89          | 0.86     |
| Gastroenterology 1/2 | 0.78                | 0.63       | 0.82          | 0.86     |
| Gastroenterology 1/4 | 0.61                | 0.45       | 0.8           | 0.67     |
| Gastroenterology 1/8 | 0.52                | 0.39       | 0.67          | 0.63     |
| Pulmonology 1/2      | 0.84                | 0.73       | 0.89          | 0.86     |
| Pulmonology 1/4      | 0.87                | 0.76       | 0.89          | 0.86     |
| Pulmonology 1/8      | 0.72                | 0.67       | 0.83          | 0.76     |
| Nephrology 1/2       | 0.8                  | 0.72       | 0.89          | 0.82     |
| Nephrology 1/4       | 0.83                | 0.63       | 0.8           | 0.8      |
| Nephrology 1/8       | 0.69                | 0.63       | 0.76          | 0.69     |
| Neurology 1/2        | 0.8                  | 0.76       | 0.82          | 0.8      |
| Neurology 1/4        | 0.8                  | 0.72       | 0.85          | 0.8      |
| Neurology 1/8        | 0.74                | 0.7        | 0.74          | 0.74     |

Table 1 says that the model accuracies for Gastroenterology 1/8, Nephrology 1/8, and Neurology 1/8 data sets prominently decreased compared with the model accuracies for the original data sets. However, the model accuracies for Pulmonology 1/8 data sets do not change too much. This shows that the size of data sets cannot be an effective measure for the uniformity of data set distribution.

This work proposed to use UMAP as a measure for the uniformity of data set distribution. As mentioned before, UMAP algorithm is very effective to visualization as well as dimension reduction and it concerns the distance between observations. Thus, UMAP effectively visualizes the distance between observations. Through this visualized distance, the uniformity of data distribution can be evaluated in a direct and intuitive manner.

As shown in Figure 1 (a), the purple dots on UMAP corresponding to the gastroenterology label gathered together on the fourth quadrant differently from other different-coloured dots. As shown in Figure 1 (b), however, the green dots on UMAP corresponding to the pulmonology label are located and distributed similarly to other different-coloured dots. As we can know from the accuracy data shown in Table 2, while the accuracy of the machine learning model (i.e., random forest model) trained with Gastroenterology 1/8 data set (corresponding to Figure 1 (a)) decreased prominently, the model accuracy for Pulmonology 1/8 data set (corresponding to Figure 1 (b)) changed just a little. This means that, if a data set distribution is ‘uniform’ in view of UMAP, the distribution does not severely
decrease the accuracy of a machine learning model. On the other hand, if a data set distribution is not ‘uniform’ in view of UMAP, the distribution does severely decrease the accuracy of a machine learning model. As shown in Figure 2, the blue dots on UMAP corresponding to the neurology label do not appear on the centre of the UMAP while the different-coloured dots are located there. Thus it can be said that the data set is not uniformly distributed in view of UMAP. This explains the reason why the corresponding accuracy decreased prominently.

**Figure 1.** UMAP for ‘Gastroenterology 1/8’ (a) vs. UMAP for ‘Pulmonology 1/8’ (b).

**Figure 2.** UMAP for ‘Neurology 1/8’.

In summary, the experiment performed in this work showed that the uniformity of the data set distribution positively affects the accuracy of the machine learning models trained by the data sets. However, the uniformity in view of data set size is insufficient to ensure the accuracy. Therefore, an evaluation method for the UMAP based uniformity of data set distribution, which represents the data validity, is proposed in this work.
Table 2. UMAPs for the training data sets.

| Random Forest Accuracy | Original Data Set | Gastroenterology 1/2 | Gastroenterology 1/4 | Gastroenterology 1/8 |
|------------------------|-------------------|----------------------|----------------------|---------------------|
| 0.89                   | 0.82              | 0.8                  | 0.67                 |

| Random Forest Accuracy | Original Data Set | Pulmonology 1/2 | Pulmonology 1/4 | Pulmonology 1/8 |
|------------------------|-------------------|-----------------|-----------------|-----------------|
| 0.89                   | 0.89              | 0.89            | 0.83            |

| Random Forest Accuracy | Original Data Set | Nephrology 1/2 | Nephrology 1/4 | Nephrology 1/8 |
|------------------------|-------------------|----------------|----------------|----------------|
| 0.89                   | 0.89              | 0.8            | 0.76            |

| Random Forest Accuracy | Original Data Set | Neurology 1/2 | Neurology 1/4 | Neurology 1/8 |
|------------------------|-------------------|---------------|---------------|---------------|
| 0.89                   | 0.82              | 0.85          | 0.74          |

4. Conclusions
According to the experiment performed in this work, additionally, the accuracies of the machine learning models trained by the TF-IDF vectorised data sets were almost double higher than those trained by the CounterVectorizer vectorised data sets. This result showed the TF-IDF vectorization produces excellent performance in natural language processing for some data sets. As mentioned above the experiment showed that the uniformity of the data set distribution positively affects the accuracy of the machine learning models trained by the data sets. Integrating these two experiment results, it can be concluded that TF-IDF is a good vectorization method for some data sets but the uniformity of the data sets shall be carefully considered in order to acquire a sufficient accuracy of a machine learning model. This conclusion is also applied to the data sets vectorised by different vectorization method, as well as the TF-IDF vectorised data sets.
This work addressed the data validity issue in view of the data set uniformity, which positively affects accuracies of machine learning model. The experiment performed in this work showed that the uniformity of the data set size distribution is insufficient to evaluate the effect to the accuracy of the machine learning models trained by the data sets. This work proposed to use UMAP as a measure for the uniformity of data set distribution. By the proposed method, data validity can be evaluated in an intuitive manner and in early phases of data science process, such as data collection and data preprocessing phases. To evaluate the data validity through the proposed method makes it possible to avoid systematically the performance decline of machine learning models due to the lack of uniformity of data sets. Particularly this method is useful in that it can be used to prevent the performance decline of machine learning models in early phases of data science process. In addition, as mentioned before, the proposed method can supplement the comparison based methods in view that results from the training or the testing a machine learning model need not to be obtained.

Future works should be done to enhance the proposed data validity evaluation method. While UMAP, which is the kernel of the proposed method, enables us to evaluate the data set uniformity visually, it does not to estimate the uniformity quantitatively. Thus, if further works make it possible to evaluate data validity quantitatively by using the distance concept of UMAP, the proposed method would be more systematic and useful. In addition, the evaluation method has been applied only to numerical data in this work. The method should be expanded to be applied to the different kinds of data such as picture data, text data, and so forth.

Acknowledgments
This paper was supported by Joongbu University Research & Development Fund in 2020.

References
[1] Howison J, Wiggins A, and Crowston K 2011 Validity issues in the use of social network analysis with digital trace data J. Assoc. Infor. Sys. 12 767
[2] Straub D, Boudreau M, and Gefen D 2004 Validation guidelines for IS positivist research Comm. Assoc. Infor. Sys. 13 380-427
[3] Son H 2019 Validity Evaluation for the Data Used for Artificial Intelligence System Proc. Intelligent Systems Conference (London) (Amsterdam: Elsevier)
[4] Becht E, Dutertre C, Kwok I, Ng L, Ginhoux F, and Newell E 2018 Evaluation of umap as an alternative to t-sne for single-cell data, bioRxiv
[5] Belkin M and Niyogi P 2002 Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neu. Infor. Proc. Sys. 2002 585–591
[6] Belkin M and Niyogi P 2003 Laplacian eigenmaps for dimensionality reduction and data representation, Neu. Comp. 15(6) 1373–1396
[7] Spivak D 2012 Metric realization of fuzzy simplicial sets Self-published Notes
[8] Laurens M 2014 Accelerating t-sne using tree-based algorithms J. Mach. Learn. Res. 15(1) 3221–3245
[9] Laurens M and Hinton G 2008 Visualizing data using t-sne J. Mach. Learn. Res. 9 2579–2605
[10] Brownlee J, 2020 How to Encode Text Data for Machine Learning with Scikit-learn Available at: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/, June 28 2020
[11] https://www.guru99.com/machine-learning-tutorial.html, 2020
[12] https://en.wikipedia.org/wiki/Ensemble_learning, July 13 2020