Karan Sarkar

Project

Executive Summary

The federal government offers many services to people with disabilities. One of these is subsidized healthcare costs. As a result, it is in the interest of insurance companies to prove to the government that some of their high cost members are in fact disabled. Insurance companies will be partially reimbursed by federal government for the medical expenses of these disabled members. In reality, the identification of disabled members is quite difficult. Current state of the art models use Tree-Bagger a machine learning model to obtain screen-in rates of around 30% for high confidence members. In this project, we will be exploring natural language processing based methods for disability identification to try and achieve higher screen-in rates.

One problem with current approaches is the so called sparsity problem. There are 70,000 distinct diagnosis codes. However, many of these codes are extremely rare and many codes are very similar. For example, there are codes for breast cancer of the left breast as well as breast cancer of the right breast. However, the current methods treat all diagnosis codes as distinct through the usage of one-hot encoded vectors. This means that we cannot generalize from one diagnosis to several closely related diagnoses. We propose generating a semantic embedding for each diagnosis based on the given text description of the diagnosis. In semantic space, similar diagnoses will have similar representations. A sequence of diagnosis vectors can then be used as input to an LSTM model that predicts disability approval.

Goal

Our goal is to test ways of leveraging the natural language representation of diagnosis codes to create better predictions of disability. We hope this new approach will yield better results than the tree based models that are used currently. We believe that generating semantic representations of diagnoses based on their text descriptions will allow better generalization than the current practice of using one-hot encoded vectors. Thus, another major goal of our project is the creation of robust diagnosis vectors. These vectors could be used in many other applications that use diagnosis codes as input.

Background and Motivation

Because of federal subsidies and disability benefits, identifying members as disabled provides a great deal of savings to insurance companies. However, it is very expensive to contact every member out of populations in the millions. Thus, the ability to provide educated guesses as to which members are disabled is very useful. As a result, improving the accuracy of these disability recognition algorithms reduces costs because fewer members need to be contacted.

The current models used are tree-based and use one-hot encoded vectors of diagnoses as input. However, because of memory and processing power constraints only about 2,000 codes out of 70,000 total are considered. These models yield a screen-in rate of about 30%. We propose that the accuracy is low because of this information loss. For example there are many codes related to diabetes mellitus. If the model uses only one diagnosis code for diabetes mellitus, it will be unable to recognize other related codes.

e0869	diabetes mellitus due to underlying condition with other specified complication
e088	diabetes mellitus due to underlying condition with unspecified complications
e089	diabetes mellitus due to underlying condition without complications
e0900	drug or chemical induced diabetes mellitus with hyperosmolarity without nonketotic hyperglycemic-hyperosmolar coma (nkhhc)
e0901	drug or chemical induced diabetes mellitus with hyperosmolarity with coma
e0910	drug or chemical induced diabetes mellitus with ketoacidosis without coma
e0911	drug or chemical induced diabetes mellitus with ketoacidosis with coma
e0921	drug or chemical induced diabetes mellitus with diabetic nephropathy
e0922	drug or chemical induced diabetes mellitus with diabetic chronic kidney disease
e0929	drug or chemical induced diabetes mellitus with other diabetic kidney complication
e09311	drug or chemical induced diabetes mellitus with unspecified diabetic retinopathy with macular edema
e09319	drug or chemical induced diabetes mellitus with unspecified diabetic retinopathy without macular edema
e093211	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy with macular edema, right eye
e093212	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy with macular edema, left eye
e093213	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy with macular edema, bilateral
e093219	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy with macular edema, unspecified eye
e093291	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, right eye
e093292	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, left eye
e093293	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, bilateral
e093299	drug or chemical induced diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, unspecified eye
e093311	drug or chemical induced diabetes mellitus with moderate nonproliferative diabetic retinopathy with macular edema, right eye
e093312	drug or chemical induced diabetes mellitus with moderate nonproliferative diabetic retinopathy with macular edema, left eye
e093313	drug or chemical induced diabetes mellitus with moderate nonproliferative diabetic retinopathy with macular edema, bilateral
e093319	drug or chemical induced diabetes mellitus with moderate nonproliferative diabetic retinopathy with macular edema, unspecified eye

Our project intends to alleviate this problem by utilizing the text descriptions of the diagnosis codes. Similar diagnosis codes have similar descriptions. We will use these text descriptions to generate dense semantic representations of the diagnoses. We hope that now all the different codes pertaining to diabetes mellitus will have similar diagnosis vectors. Thus, we will now be able to generalize to related diagnoses.

Related Papers

The paper ‘What can natural language processing do for clinical decision support?’ surveys some approaches for applying natural language processing to making medical decisions. Dina Demner-Fushman, Wendy W. Chapman and Clement J.McDonald look at how models such as Hidden Markov Models and Maximum Entropy can be used to predict medical outcomes. They find the the wordings work in general better than codes in making decisions in areas such as radiology.

Approaches

Our basic approach is to generate a vector to represent each member. Each member vector will then be fed to a shallow neural network classifier. However, we will have two approaches for generating the member vectors

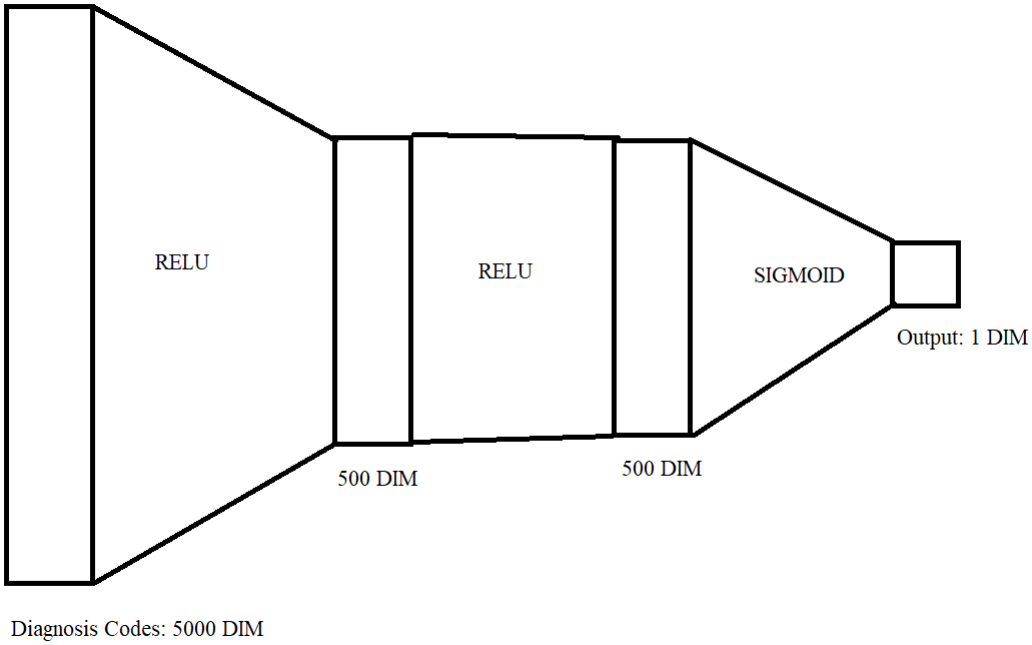
A. The diagnoses are truncated and then one-hot encoded. This is the baseline control group method that does not use any natural language processing.
B. We generate embeddings for each diagnosis based on using an LSTM to predict procedure codes. This is the experimental group that uses natural language processing.

In approach A, we handle the sparsity problem by truncating the codes. As noted above, similar diagnoses have similar codes as with the diabetes example. However, truncating the codes can leave out certain important details. For example, cancer in remission is normally grouped together with cancer not in remission. This can cause problems, because whether or not someone is in remission may affect their chanced at being approved for disability benefits.

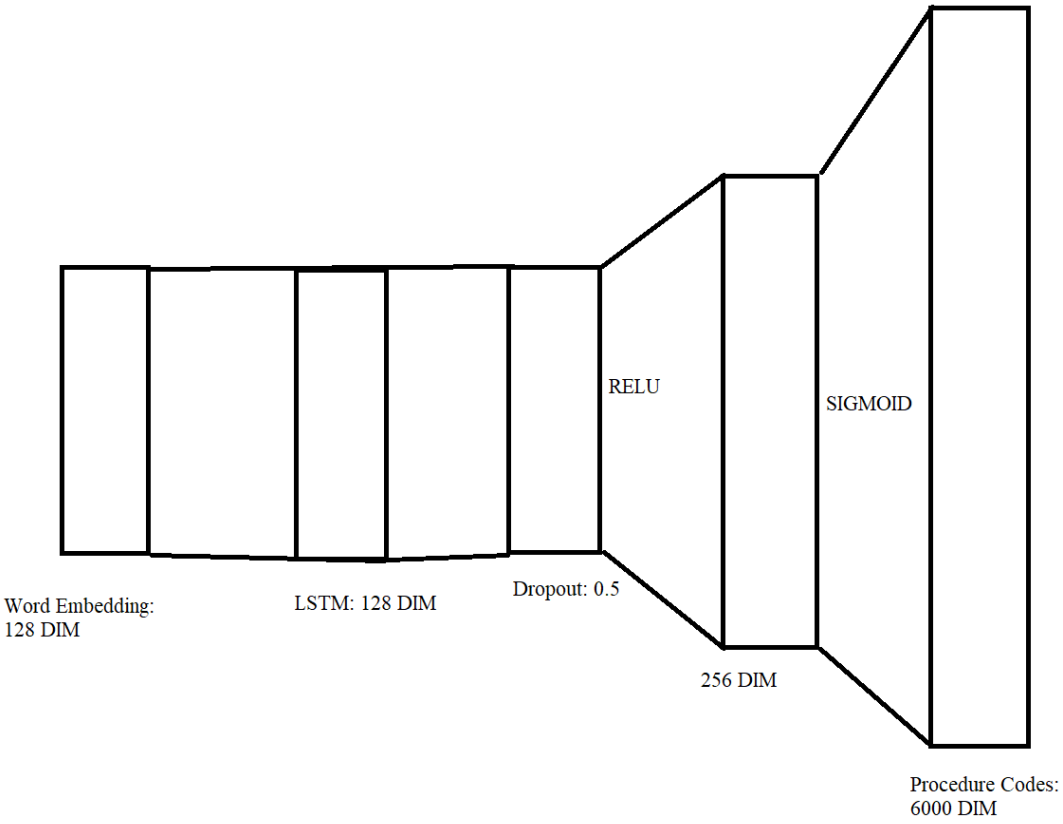
Our goal is to observe whether using natural language processing yields an improvement in accuracy. We believe that Approach B will allow individual words to influence the representation of a diagnosis. In theory, now the phrases ‘in remission’ and ‘not in remission’ will be effectively weeded out. Essentially, now the sparsity problem is being solved by using natural language processing.

System Architecture

Approach A:



Approach B:



In approach B, for each diagnosis code, we compute a procedure code vector that acts like a diagnosis code embedding. Then, each of the procedure code vectors is summed to create a member vector. The member vector is then fed to a classifier that is identical to the one in approach A.

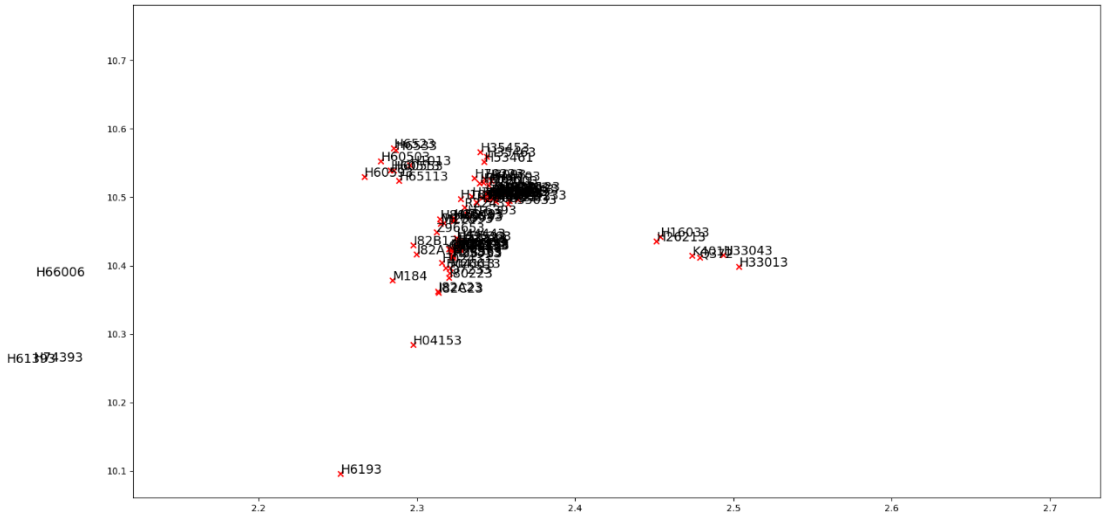
LSTM Details:

Our neural network required a fixed sequence length. We used the length of the longest description which was about 20. To make all descriptions fit this length, we padded with blanks to the front of the sequence. This way the LSTM would see the meaningful words towards the end.

We handled out of vocabulary words by setting them to zero vectors. Because we did not have many out of vocabulary words, this was not really a major problem. The vocabulary was turned into vectors using pre-trained Pub-med embeddings we obtained from Professor Ji.

Visualization

We can use the TSNE dimensionality reduction algorithm to visualize the embeddings in two-dimensional space. We can do this to see for sure whether the embedding is meaningful.



From this diagram, we can see that similar diagnoses are being clustered together. In this particular case, all of these diagnosis codes correspond to ear infections. This can be easily noted because all of the codes begin with ‘H’. Because they have both similar descriptions and procedures, they are clustered together. This ensures that our algorithm generalizes between similar diagnoses.

Results

We trained each neural network for about 5 epochs until they began to overfit. Approach A got an accuracy of about 90%. Approach B got an accuracy of 93%. Thus using natural language processing to handle the diagnoses was an improvement over the previous methods. We conclude that the order and phrasing of the descriptions is indeed very important in determining disability.

Future Work

Given the success of diagnosis descriptions in predicting disability, we think that additional natural language features such as doctor’s notes and procedure descriptions may be of use as well. This could be used to improve performance into the future. Hopefully, there might emerge some sort of universal embedding that can be added to by a number of different document and data types. This would solve one of the problems in medical classification today which is the existence of multiple data types and inconsistency with the availability of these data types for different members.

Sources

J. A. Perez-Ortiz and M. L. Forcada, "Part-of-speech tagging with recurrent neural networks," IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), Washington, DC, USA, 2001, pp. 1588-1592 vol.3.