2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Subsumption reduces dataset dimensionality without decreasing performance of a machine learning classifier

Donald C. Wunsch III[1] and Daniel B Hier[1]

Abstract— **When features in a high dimension dataset are organized hierarchically, there is an inherent opportunity to reduce dimensionality. Since more specific concepts are subsumed by more general concepts, subsumption can be applied successively to reduce dimensionality. We tested whether subsumption could reduce the dimensionality of a disease dataset without impairing classification accuracy. We started with a dataset that had 168 neurological patients, 14 diagnoses, and 293 unique features. We applied subsumption repeatedly to create eight successively smaller datasets, ranging from 293 dimensions in the largest dataset to 11 dimensions in the smallest dataset. We tested a MLP classifier on all eight datasets. Precision, recall, accuracy, and validation declined only at the lowest dimensionality. Our preliminary results suggest that when features in a high dimension dataset are derived from a hierarchical ontology, subsumption is a viable strategy to reduce dimensionality.**

Clinical relevance— **Datasets derived from electronic health records are often of high dimensionality. If features in the dataset are based on concepts from a hierarchical ontology, subsumption can reduce dimensionality.**

I. Introduction and Previous Work

Electronic health records (EHR) hold huge amounts of clinical data. Some of the value of this data can be unlocked by machine learning [1], [2]. It is estimated that the EHR system of a large healthcare organization holds clinical information equivalent to 100 million years of patient data (10 million patients times 10 years) [3]. Each hospital encounter generates as much as 150,000 pieces of data. Although some hospital data is numerical (e.g. laboratory results), admission notes, progress notes, and discharge summaries are difficult to convert to a computable form. One approach to making the *signs and symptoms* of patients computable has been called *deep phenotyping*. With deep phenotyping, the signs and symptoms of patients are represented as concepts from an ontology such as the Human Phenotype Ontology (HPO) [4]–[6].

Disease classification is an important goal of machine learning healthcare applications [1]. The signs and symptoms of patients are important features utilized by machine learning classifiers to make medical diagnoses. Healthcare datasets are generally of high dimensionality with hundreds or thousands of features. For example, the Human Phenotype Ontology, used to encode the signs and symptoms of subjects with human diseases, has 19,249 unique concepts, offering "a standardized set of phenotypic terms that are organized in a hierarchical fashion. Using standardized hierarchies enables us to put our phenotypic knowledge into an organized framework that can be analyzed by computational means" [7].

Feature selection (dimension reduction) is important to machine learning applications, especially for datasets of high dimensionality. Feature selection can improve model accuracy, reduce over-fitting, eliminate irrelevant features, reduce computation costs, and improve model interpretability [8], [9]. Approaches to reducing feature dimensionality have included filter methods, wrapper methods, ensemble methods, principal components analysis, and genetic algorithms [8]–[10].

Ontologies offer a unique additional opportunity for dimension reduction due to their inherent hierarchical structure. Most medical terminology ontologies are based on a *subsumptive containment hierarchy* with classes hierarchically organized from the general to the specific; also known as *IS-A hierarchies.* Each child class inherit properties from its parent class. The inheritance of properties from a parent is called *subsumption*. Subsumption supports dimension reduction. For example, the children concepts micrographia, masked face, impaired turns, decreased arms swing, reduced blink rate are subsumed under the more general concept *bradykinesia* (Fig. 1). Similarly the concepts fine tremor, resting tremor, action tremor, postural tremor, voice tremor, senile tremor are subsumed under the more general concept *tremor*. The hierarchical structure of ontologies and the ability to collapse sub-classes into more general super-classes makes an ontology well-suited for feature reduction.

II. Methods

A. Proposed Approach

We proposed to test the hypothesis that the hierarchical structure of ontologies can be used to reduce the dimensionality of disease datasets without an adverse impact classification accuracy. We tested this hypothesis on a disease dataset with 168 instances (patients), 293 unique features (signs and symptoms), 1953 total features, and 14 unique labels (diagnoses). The dataset was derived from published case histories in neurology textbooks as previously described [11] and no protected health information (PHI) was used in this study. Features were derived from a hierarchical ontology with 1242 unique concepts based on the neurological examination [12], [13]. We tested classification accuracy, precision, and recall at 8 different levels of specificity within the ontology hierarchy, reflecting a reduction in dataset dimensionality from 293 to 11 dimensions.

[1]Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla MO 65401, USA dcwq46.mst.edu, hierd.mst.edu

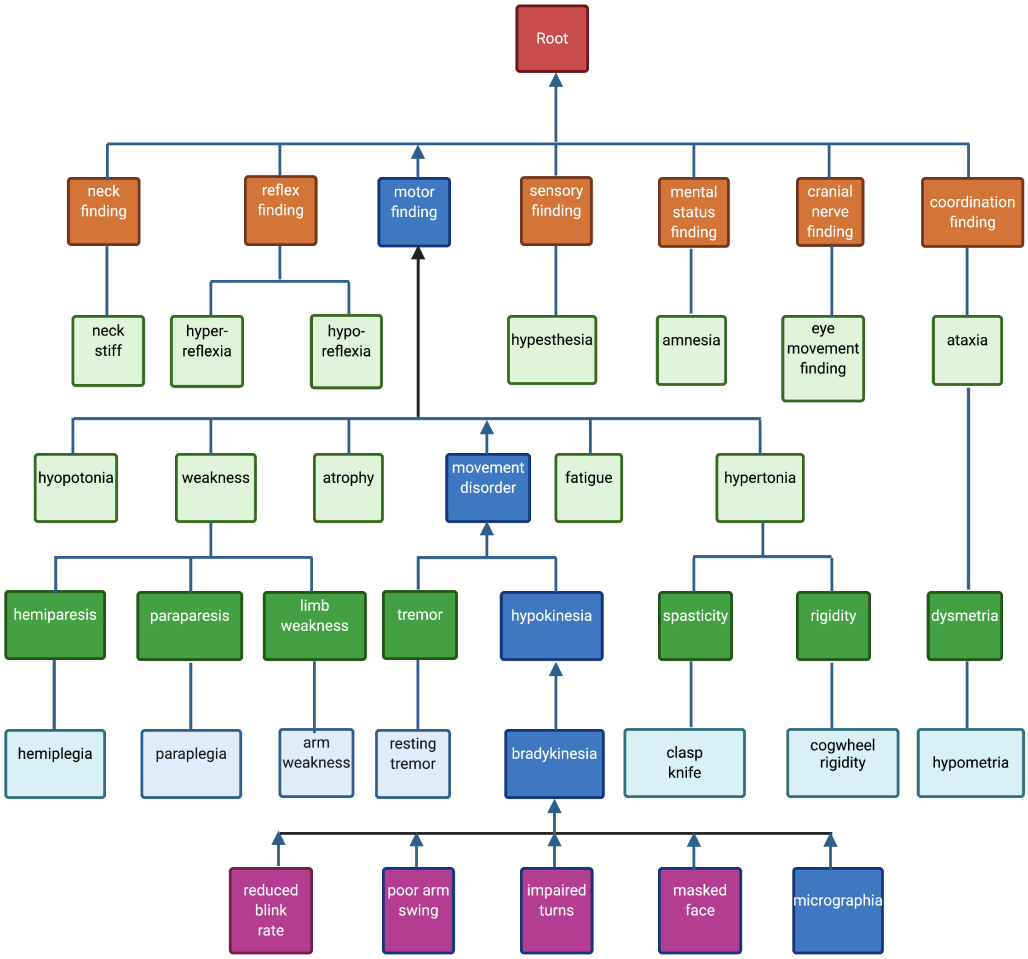

Fig. 1. A small excerpt from the neuro-ontology. The neuro-ontology has 11 major branches below the root (seven shown) and 1242 terminal nodes. Concepts in the ontology become increasingly specific at lower levels going from coarsest (least specific) to most granular (most specific) at the lowest level. The concept *micrographia* (shown in dark blue) is most specific and is subsumed by *bradykinesia*, then *movement disorder*, and finally by the coarsest (least specific) concept *motor finding*. Each color represents a different level in the concept hierarchy.

TABLE I

DIAGNOSES AND TYPICAL FINDINGS

Diagnosis	N	Finding
amyotrophic lateral sclerosis	22	weakness, fasciculations, hyperreflexia
dystonia	10	dystonia
normal pressure hydrocephalus	14	dementia, gait apraxia, incontinence
Lewy body dementia	6	dementia, hallucinations, bradykinesia
hemiballismus	4	hemiballismus
myasthenia gravis	18	weakness, diplopia, ptosis
moyopathy	18	proximal weakness
Huntington disease	17	personality change, chorea, dementia
essential tremor	7	tremor
Parkinson disease	20	tremor, bradykinesia, rigidity
multiple system atrophy	9	dysautonomia, bradykinesia, rigidity
progressive supranuclear palsy	9	gaze palsy, bradykinesia, rigidity
spinocerebellar ataxia	5	ataxia, weakness, spasticity
Wilson disease	9	tremor, ataxia, personality change

B. Dimensionality reduction

We used Python to traverse the neuro-ontology [12]from each of its 1242 terminal nodes to the root node (Fig. 1). We created 1242 ordered lists (one for each concept) of length n=8 where the last element in the list was the penultimate concept (last node prior to root) and the first element in the list was the terminal concept. If the list was less than 8 elements long, it was back-filled to 8 elements by repeating the first element (terminal node) until all lists were 8 elements in length. For example the list for *micrographia* (Fig. 1) was *[micrographia, micrographia, micrographia, micrographia, bradykinesia, hypokinesia, movement disorder, motor finding]*. Using these ordered lists as a reference, we created eight new datasets by sequentially replacing the first element in the ordered list with the second element and so on, seven times. This allowed us to perform dimension reduction sequentially with each reduction reflecting replacement of a child concept with its parent concept (Table II.)

C. Machine learning classifier and classification metrics

We used MATLAB to construct a multilayer perceptron (MLP) of 3 hidden layers, each with 300 neurons. Each neuron utilized a hyperbolic tangent transfer function. Output layers used a softmax transfer function. The learning rate was set at 0.01 with a momentum constant of 0.1. Our dataset

was split into training, testing, and validation subsets using a 70:15:15 ratio respectively. Each trial was constrained to a maximum of 1000 epochs (most trials ran for fewer than 60 epochs). Training performance was evaluated by cross-entropy, which consistently yielded higher classification accuracy than a mean-squared error performance metric [14].

Each classification was one-against-rest (OAR). The limited size of the dataset precluded meaningful classification results with some of the diagnosis classes with few members (Table I). Accuracy, precision, recall, and minimum validation loss were recorded and averaged across 10 trials at each of the eight ontology levels. Two-way ANOVA and post hoc testing were by GraphPad Prism 9.

TABLE II
DIMENSIONALITY

Level	Features
level 1	293
level 2	287
level 3	272
level 4	255
level 5	222
level 6	157
level 7	62
level 8	11

III. Results

A. Dimension reduction

Using sequentially repeated subsumption based on hierarchical levels in the ontology, we reduced dimensionality from 293 dimensions to 11 dimensions (Table II). Each case was represented by eight different vectors of successively lower dimensionality based on the hierarchy of signs and symptoms in the neuro-ontology.

B. Classification performance

We tested the MLP classifier on the four most common diagnoses in the dataset (amyotrophic lateral sclerosis, myopathy, myasthenia gravis, and Parkinson disease (Table I). Classification precision, accuracy, recall, and validation loss did not decline until level 8 (the level that utilized the most general concepts) of the ontology (Figs. 2-5). In general, the classifier performed well on all four diagnoses. Classification performance was minimally better for the diagnosis of myasthenia gravis than the other three diagnoses (Figs. 2-4).

IV. Discussion and Conclusions

Like many disease datasets, our dataset was of high dimensionality (293 different signs and symptoms) despite having only 168 cases (Table I). The features of our dataset were derived from a subsumptive containment hierarchy [12]. In a subsumptive containment hierarchy more specific concepts are subsumed by more general concepts. We used subsumption successively to reduce the dimensionality of our dataset from 293 dimensions to 11 dimensions. Each successive application of subsumption reduced dimensionality of the dataset and substituted a more general concepts for a more specific concepts. The performance of the MLP classifier was surprisingly lossless with dimension reduction.

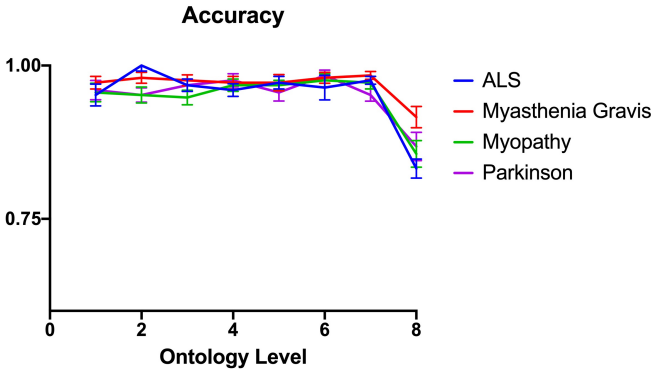

Fig. 2. Accuracy $(mean \pm SEM)$ for classification by ontology level. Two-way ANOVA showed significant effects ($p < 0.05$) for both ontology level and diagnosis. Post hoc tests (Tukey) showed level 8 accuracy was lower than other levels and that myasthenia gravis accuracy was higher than Parkinson disease and myopathy ($p < 0.05$).

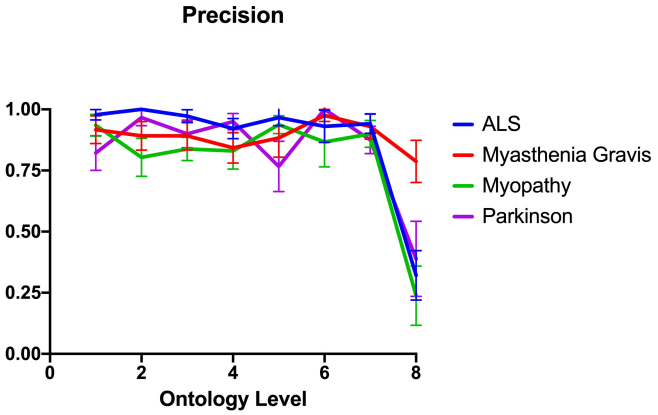

Fig. 3. Precision $(mean \pm SEM)$ by ontology level. Two-way ANOVA showed significant effects ($p < 0.05$) for both ontology level and diagnosis. Post hoc tests (Tukey) showed level 8 precision lower than the other levels and myasthenia gravis precision higher than myopathy.

Performance of the classifier did not drop significantly until the eighth level of the ontology which utilized the most general concepts. At the seventh level of the ontology, dimensionality was reduced to 62 dimensions from 293 dimensions (a 79% reduction), yet overall performance of the classifier remained high (Figs 2-5).

The goal of dimension reduction methods for high dimension datasets is to find the minimal subset of features that maintains classifer accuracy and retains predicted class sizes reflective of the class sizes in the ground truth dataset upon retraining [15]–[17]. Two commonly used strategies to reduce dataset dimensionality include feature selection and feature extraction. Feature selection (filter methods, wrapper methods) emphasize algorithms that reduce the number of features into the smallest subset that accurately predict class membership [15]–[17]. Feature extraction methods (principal components, linear discriminant analysis, etc.) emphasize methods for collapsing a large number of features into a smaller number of highly predictive features. The use of subsumption to collapse features into a smaller number of

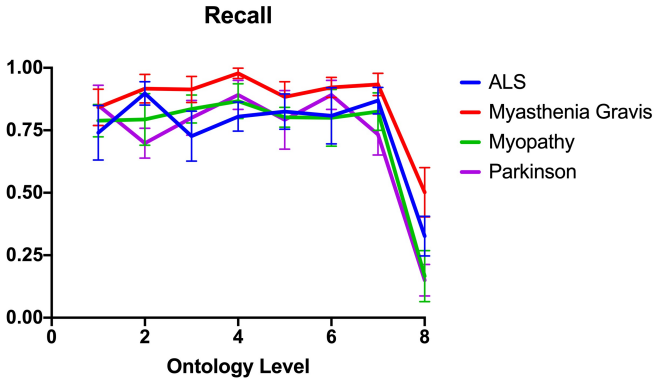

Fig. 4. Recall $(mean \pm SEM)$ by ontology level. Two way ANOVA showed both ontology level (df= 7) and diagnosis (df=3) effects were significant (p <0.05). Post hoc testing with Tukey correction showed ontology level 8 had lower recall than the other 7 levels. Recall was better for myasthenia gravis (p<0.05) than the other three diagnoses.

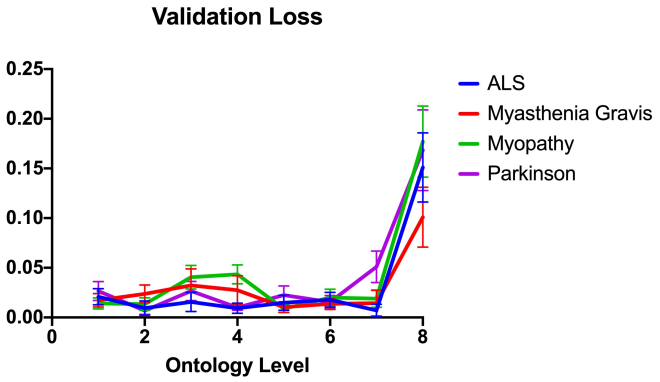

Fig. 5. Validation loss $(mean \pm SEM)$ by ontology level. Two-way ANOVA showed ontology level was significant ($p < 0.05$). Diagnosis effect was non-significant. Post hoc comparisons with Tukey correction showed level 8 validation loss was higher than other levels ($P < 0.05$).

features bears more resemblance to a feature extraction strategy than a feature selection strategy. The use of knowledge embedded in a hierarchical ontology has been suggested by others as a dimension reduction strategy [18].

This work has important limitations. First, the dataset was small and future testing utilizing a larger dataset will be advantageous. Second, we did not test our dataset on other classifiers such as SVM, k-nearest neighbor, or logistic regression [19]. Comparison of the MLP classifier to other classifiers would be instructive. Third, due to asymmetries in the depth of the ontology, significant dimension reduction did not occur until level 5 of the ontology (Table II). Finally, we did not compare subsumption to other feature selections methods such as FCBF [20], mutual information [21], or Relief [22]. We plan to make these comparisons in the future. Other studies have found that when different feature reduction strategies are compared classifier performance depends on the nature of the dataset, the classifer utilized, as well as the feature reduction algorithm [19].

REFERENCES

[1] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[3] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[4] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott *et al.*, "The human phenotype ontology in 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D865–D876, 2017.

[5] S. Köhler, N. C. Øien, O. J. Buske, T. Groza, J. O. Jacobsen, C. McNamara, N. Vasilevsky, L. C. Carmody, J. Gourdine, M. Gargano *et al.*, "Encoding clinical data with the human phenotype ontology for computational differential diagnostics," *Current protocols in human genetics*, vol. 103, no. 1, p. e92, 2019.

[6] T. Groza, S. Köhler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel, L. M. Schriml, W. A. Kibbe, P. N. Schofield, T. Beck *et al.*, "The human phenotype ontology: semantic unification of common and rare disease," *The American Journal of Human Genetics*, vol. 97, no. 1, pp. 111–124, 2015.

[7] The National Center for Biomedical Ontology, "The human phenotype ontology," https://bioportal.bioontology.org/ontologies/HP, 2021, uploaded: 2020-12-07.

[8] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[9] M. Kuhn and K. Johnson, *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[10] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in *2014 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2014, pp. 1–6.

[11] D. B. Hier, J. Kopel, S. U. Brint, D. C. Wunsch, G. R. Olbricht, S. Azizi, and B. Allen, "Evaluation of standard and semantically-augmented distance metrics for neurology patients," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–15, 2020.

[12] D. B. Hier and S. U. Brint, "A neuro-ontology for the neurological examination," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–9, 2020.

[13] NCBO BioPortal, "Neurologic examination ontology," \https://bioportal.bioontology.org/ontologies/NEO, Accessed: 2021-01-05.

[14] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[15] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[16] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[17] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[18] D. C. Corrales, E. Lasso, A. Ledezma, and J. C. Corrales, "Feature selection for classification tasks: Expert knowledge or traditional methods?" *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 2825–2835, 2018.

[19] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *New challenges for feature selection in data mining and knowledge discovery*, 2008, pp. 90–105.

[20] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

[21] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[22] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of biomedical informatics*, vol. 85, pp. 189–203, 2018.