MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 5, Issue 1, January-June 2020

Self-Supervised Representation Learning for Structured Data

Surender R Pathak
Assistant Professor
Department of Artificial Intelligence
Vidya Jyoti Technical Campus
Email: *Surrender_r_pathak@gmail.com*

Abstract

Structured data such as tables, graphs, and relational records constitute the backbone of industrial analytics, healthcare informatics, finance, and engineering monitoring systems. However, learning meaningful representations from structured data remains challenging due to heterogeneity, sparsity, and limited labeled samples. Self-supervised representation learning has recently emerged as a promising paradigm to extract informative latent features without heavy reliance on manual annotations. Instead of labels, the learning objective is derived from intrinsic structure of data such as attribute relations, topology, or contextual dependencies. This paper reviews major self-supervised learning (SSL) approaches for structured data including tabular data, graph-structured data, relational databases, and time-series tables. We discuss pretext task design, contrastive and generative frameworks, and hybrid architectures combining neural networks with statistical models. Comparative analysis of methods is provided along with application domains such as anomaly detection, recommender systems, predictive maintenance, and healthcare risk modeling. The study also outlines challenges including feature heterogeneity, missing values, scalability, and evaluation metrics. Finally, emerging directions such as multimodal structured SSL and foundation models for tabular learning are discussed. The review aims to provide a consolidated understanding of self-supervised representation learning for structured datasets and its role in future intelligent data systems.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

Keywords: *Self-supervised learning, structured data, representation learning, tabular data, graph learning, contrastive learning, pretext tasks*

INTRODUCTION

Most real-world data used in decision systems is structured rather than unstructured. Examples include relational tables in enterprise databases, sensor logs in industrial plants, financial transaction records, and knowledge graphs. Unlike images or text, structured data has heterogeneous feature types (categorical, ordinal, numerical), relational dependencies, and irregular distributions. Traditional machine learning approaches rely on feature engineering and labeled data, which are costly and domain dependent.

Self-supervised learning (SSL) addresses this limitation by learning representations from unlabeled data through surrogate objectives known as pretext tasks. The core idea is that meaningful supervision signals can be derived from inherent structure within data itself. SSL has achieved remarkable success in computer vision and natural language processing, but its adaptation to structured data is still evolving.

Structured data differs significantly from perceptual data. Relationships among features, table rows, or graph nodes carry semantic meaning. For instance, correlations between attributes in a medical table or connectivity patterns in a power grid network encode important knowledge. SSL methods must therefore capture both feature-level and relational structure.

This paper reviews state-of-the-art self-supervised representation learning techniques for structured data. We categorize methods by data type and learning objective, analyze strengths and limitations, and discuss future directions for large-scale structured SSL.

FUNDAMENTALS OF SELF-SUPERVISED LEARNING

Self-supervised learning (SSL) is a representation learning paradigm in which supervision signals are derived automatically from the intrinsic structure of data rather than from manually annotated labels. Instead of relying on external ground truth, the model formulates auxiliary learning objectives that require understanding relationships among features, samples, or structural components. Through solving these surrogate tasks, the model learns latent representations that capture semantic, statistical, or structural regularities present in the dataset.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

In structured data contexts, SSL is particularly valuable because labeled datasets are often scarce or expensive to obtain. Industrial logs, financial records, healthcare tables, and relational databases contain abundant unlabeled information but limited annotations. SSL enables leveraging this large unlabeled corpus to learn meaningful embeddings that can later support supervised or unsupervised downstream tasks.

From a theoretical viewpoint, SSL can be interpreted as learning invariances and dependencies within data. For structured datasets, these dependencies may include attribute correlations in tables, relational links across entities, graph topology, or temporal continuity. The learned embedding vector therefore acts as a compact representation of these relationships, facilitating improved generalization and robustness in predictive models.

1. Pretext and Downstream Tasks

Self-supervised learning typically operates in two stages: a pretext learning phase and a downstream utilization phase. This separation allows models to first acquire general structural knowledge from unlabeled data and then adapt it to specific application objectives.

- **Pretext task:**

A surrogate learning objective constructed from inherent data structure. The model is trained on unlabeled data to solve this task, thereby learning internal representations.

- **Downstream task:**

A practical machine learning objective such as classification, regression, clustering, anomaly detection, or forecasting that uses the learned representations as features.

The success of SSL depends strongly on the design of the pretext task. A well-designed pretext task should satisfy three conditions:

- It must require understanding meaningful relationships in data
- It should avoid trivial shortcuts
- It should align with downstream objectives

For structured data, pretext tasks often exploit feature relations, row similarity, topology, or temporal ordering. Below are common examples with elaboration.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

Predict Masked Table Values

In this task, selected attributes within a row are masked and the model predicts their values using remaining attributes. This forces the encoder to learn inter-feature dependencies and correlations.

For example, in a healthcare table, predicting a masked "disease risk score" from demographics and lab values requires understanding medical relationships. Similarly, predicting a missing sensor measurement from other sensor channels captures system dynamics.

This task is particularly useful because:

- It naturally handles missing data scenarios
- It models conditional feature distributions
- It encourages holistic row representation

Predict Relationship between Rows

Rows in structured datasets may exhibit similarity or relational links (e.g., same user, same machine, and same location). The model learns whether two rows are related or similar.

Positive pairs may include:

- Two records of same entity
- Similar feature patterns
- Temporal neighbors

Negative pairs are unrelated rows. By distinguishing them, the model learns similarity-preserving embeddings that reflect entity or pattern identity.

This task supports applications such as clustering, recommendation, and anomaly detection.

Reconstruct Missing Attributes

Unlike masked modeling where few attributes are hidden, reconstruction tasks may involve predicting multiple missing features or even entire subsets of a row. This is often implemented using autoencoders or denoising networks.

The model compresses input into latent representation and reconstructs original data. To succeed, the representation must capture joint distribution of features. This encourages learning of global structure rather than local correlations only.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

Predict Graph Neighbors

For graph-structured data, nodes are connected through edges representing relations. A self-supervised task can require predicting neighbors of a node or determining whether two nodes are connected.

This objective encourages embeddings that encode graph topology. Nodes with similar neighborhoods become close in representation space, reflecting structural equivalence or community membership.

Such representations are useful for link prediction, recommendation, and graph clustering.

Temporal Order Prediction

In temporal structured datasets, records follow chronological sequence. A pretext task can require predicting the correct order of shuffled time segments or determining whether a sequence is in correct temporal direction.

To solve this, the model must learn temporal dependencies and dynamics. This helps capture seasonality, trends, and causal progression in time-series tables such as sensor logs or financial data.

Representation Transfer to Downstream Tasks

After pretext training, the learned encoder generates embeddings for each instance (row, node, or sequence). These embeddings can then be used in multiple downstream scenarios:

- Input features for classifiers or regressors
- Similarity space for clustering
- Distance measure for anomaly detection
- State representation for forecasting

Because the representation already encodes structural dependencies, downstream models require fewer labels and achieve better generalization.

2. SSL Framework Types

Self-supervised learning for structured data can be broadly categorized into three main frameworks based on the nature of the learning objective: generative, contrastive, and

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

predictive. Each framework captures different aspects of structure and offers complementary advantages.

GENERATIVE SELF-SUPERVISED LEARNING

Generative SSL aims to model the underlying data distribution by reconstructing input data from a latent representation. The encoder compresses structured input into embedding vector, and the decoder reconstructs original attributes, rows, or topology.

Common implementations include:

- Autoencoders
- Variational autoencoders
- Masked modeling networks

For structured data, generative SSL learns feature correlations and joint attribute distributions. For example, reconstructing a full row from partial attributes requires understanding how features co-occur. Similarly, reconstructing adjacency matrix in graphs captures connectivity patterns.

Advantages:

- Captures global structure and dependencies
- Handles missing or corrupted data
- Provides interpretable reconstruction error

Limitations:

- May learn trivial identity mapping
- Reconstruction objective may not align with downstream tasks
- Sensitive to noise distribution

Despite limitations, generative SSL remains effective for tabular and relational representation learning.

CONTRASTIVE SELF-SUPERVISED LEARNING

Contrastive SSL learns representations by comparing positive and negative sample pairs. The objective is to maximize similarity between related samples while minimizing similarity between unrelated samples.

In structured data, positive pairs may include:

- Augmented versions of same row
- Neighboring graph nodes

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

- Temporal segments of same entity

Negative pairs are unrelated samples. The encoder maps data into embedding space where positive pairs cluster together and negatives are separated.

Contrastive learning captures invariant features — properties that remain consistent under transformations or across related instances. For example, two augmented versions of a table row share same semantics even if some attributes are dropped or perturbed.

Advantages:

- Produces discriminative embeddings
- Robust to noise and augmentation
- Effective for clustering and similarity tasks

Limitations:

- Requires careful augmentation design
- Needs negative sampling strategy
- Computationally intensive for large datasets

Contrastive SSL has become dominant in graph representation learning and is increasingly applied to tabular data.

PREDICTIVE SELF-SUPERVISED LEARNING

Predictive SSL focuses on predicting structural or contextual properties of data rather than reconstructing raw input. The objective is to infer missing structural information such as neighbors, links, attributes, or temporal states.

Examples in structured data include:

- Predict node degree or community
- Predict next time-step value
- Predict relation type
- Predict attribute category

This framework emphasizes understanding relationships rather than exact reconstruction. It is especially suited for relational and graph data where structure itself carries meaning.

Advantages:

- Aligns closely with structural tasks
- Avoids trivial reconstruction

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

- Captures relational semantics

Limitations:

- Requires domain-specific task design
- May ignore raw feature details
- Limited generality across datasets

STRUCTURED DATA TYPES AND REPRESENTATION CHALLENGES

Structured datasets vary widely in form and complexity. Major categories include tabular, relational, graph, and temporal structured data.

1. Tabular Data

Tabular data contains rows (instances) and columns (features). Features may be numerical or categorical. Challenges include:

- Mixed feature types
- Missing values
- Feature correlations
- Non-Euclidean relationships

Traditional deep learning struggles with tabular representation because feature order is not meaningful and interactions are sparse.

2. Graph-Structured Data

Graph data represents entities and relations as nodes and edges. Examples:

- Social networks
- Biological networks
- Knowledge graphs
- Infrastructure networks

Challenges:

- Irregular connectivity
- Variable neighborhood size
- Multi-relation edges
- Scalability

Graph neural networks (GNNs) provide structural inductive bias but require labeled nodes. SSL enables unsupervised graph representation learning.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

3. Relational Databases

Relational databases consist of multiple linked tables with foreign keys. Representation must capture:

- Inter-table dependencies
- Schema structure
- Join relationships

Learning from relational databases without flattening is still an emerging area.

4. Temporal Structured Data

Time-series tables arise in sensor monitoring and finance. Representation learning must model:

- Temporal correlation
- Multivariate dependencies
- Seasonality
- Missing sequences

SSL approaches often use forecasting or masked reconstruction tasks.

SELF-SUPERVISED LEARNING FOR TABULAR DATA

Tabular SSL aims to learn feature embeddings and row representations from unlabeled tables.

1. Masked Attribute Modeling

Inspired by masked language modeling, random features are hidden and predicted from remaining attributes. This forces model to capture feature relationships.

Objective:

Predict masked column value given other columns.

Benefits:

- Captures feature dependencies
- Handles missing values
- Improves downstream prediction

2. Contrastive Row Representation Learning

Rows with similar attributes are treated as positive pairs, while random rows are negatives. Augmentations include:

- Feature dropout
- Noise injection

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

- Mixup

The encoder learns similarity-preserving embedding.

3. Autoencoder-Based Tabular SSL

Autoencoders reconstruct entire rows from compressed representation. Variants include:

- Denoising autoencoders
- Variational autoencoders
- Feature-wise autoencoders

They model joint distribution of attributes.

Table 1: Comparison of Tabular SSL Methods

Method	Pretext Task	Strength	Limitation
Masked modeling	Predict hidden features	Captures correlations	Sensitive to feature imbalance
Contrastive row SSL	Similarity learning	Robust embedding	Needs augmentation design
Autoencoder	Reconstruction	Handles noise	May ignore semantic relations
Hybrid SSL	Combined tasks	High accuracy	Higher complexity

SELF-SUPERVISED LEARNING FOR GRAPH DATA

Graph SSL is one of the most mature structured SSL areas.

1. Node-Level Contrastive Learning

Graph augmentations generate two views of same node (e.g., edge dropout, feature masking).

Model maximizes similarity between views while separating others.

Objective:

Nodes from same graph neighborhood → positive

Different nodes → negative

This preserves structural identity.

2. Graph Context Prediction

Model predicts subgraph context or neighbors of a node. This captures topology and

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

connectivity.

Examples:

- Predict node degree
- Predict neighbors
- Predict community

3. Graph Autoencoders

Graph autoencoders reconstruct adjacency matrix or node features. Latent embedding preserves structure.

Applications:

- Link prediction
- Community detection
- Fraud detection

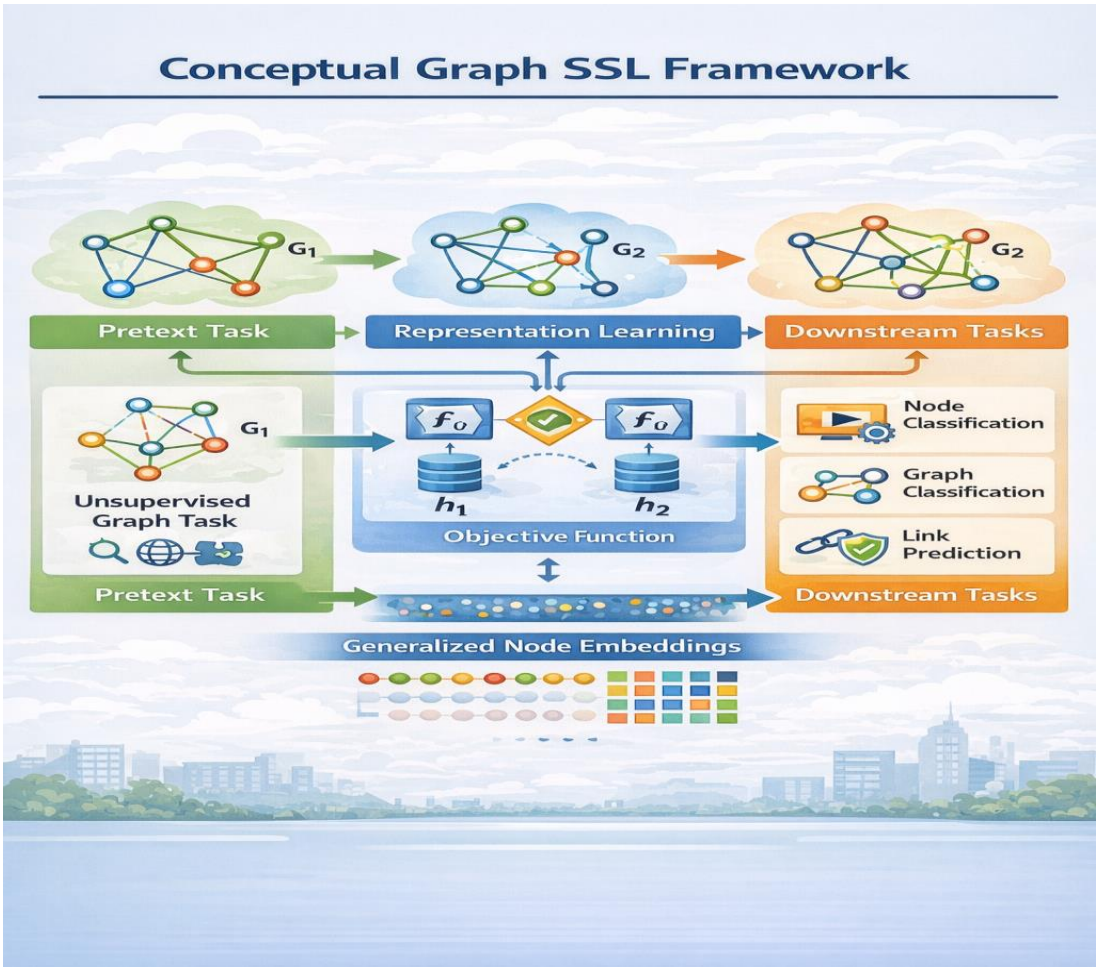


Figure 1: Conceptual Graph SSL Framework

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

SELF-SUPERVISED LEARNING FOR RELATIONAL DATA

Relational SSL extends tabular learning to multiple linked tables.

1. Schema-Aware Representation

Relational schema defines relationships among tables. SSL tasks use schema as supervision:

- Predict foreign key links
- Predict table membership
- Predict join relations

2. Cross-Table Masked Modeling

Values across linked tables are masked and predicted jointly. This captures inter-table dependencies.

Example:

Predict customer attribute from transaction table.

3. Relational Graph Conversion

Relational databases can be converted into heterogeneous graphs. SSL methods for graphs can then be applied.

Advantages:

- Unified representation
- Handles multi-relations
- Captures joins naturally

SELF-SUPERVISED LEARNING FOR TEMPORAL STRUCTURED DATA

Temporal structured data includes multivariate time-series tables.

1. Forecasting as Pretext Task

Model predicts future values from past sequence. Representation captures temporal patterns.

Applications:

- Predictive maintenance
- Demand forecasting
- Health monitoring

2. Temporal Contrastive Learning

Segments from same time-series are positive pairs; different series are negatives.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

Augmentations:

- Time warping
- Cropping
- Noise

3. Masked Time-Step Modeling

Random time points are hidden and predicted. Similar to masked modeling in sequences.

Table 2: Temporal SSL Pretext Tasks

Task	Objective	Captured Property
Forecasting	Future prediction	Temporal dependency
Masked step	Missing reconstruction	Local continuity
Segment contrast	Similarity	Global pattern
Order prediction	Sequence order	Temporal direction

APPLICATIONS OF SELF-SUPERVISED STRUCTURED REPRESENTATION

Self-supervised structured embeddings enable multiple downstream applications.

1. Anomaly Detection

In industrial or financial tables, anomalies deviate from learned representation distribution. SSL helps detect rare patterns without labels.

Examples:

- Fraud detection
- Fault detection
- Intrusion detection

2. Recommender Systems

User-item interaction tables benefit from SSL embeddings capturing similarity and preferences. Contrastive learning improves recommendation quality.

3. Healthcare Risk Modeling

Electronic health records are structured relational tables. SSL can learn patient representation from visits, diagnoses, and lab data.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

Applications:

- Disease prediction
- Patient clustering
- Treatment recommendation

4. Predictive Maintenance

Sensor tables and time-series logs represent machine behavior. SSL captures degradation patterns for failure prediction.

5. Knowledge Graph Completion

Graph SSL embeddings help infer missing relations or links.

CHALLENGES IN SELF-SUPERVISED LEARNING FOR STRUCTURED DATA

Despite progress, several challenges remain.

1. Feature Heterogeneity

Structured data contains mixed data types. Neural encoders must handle categorical and numerical features jointly.

2. Missing and Noisy Data

Tables often have missing values and measurement noise. SSL objectives may propagate errors.

3. Pretext Task Design

Unlike images, augmentations for structured data are not obvious. Incorrect augmentation may destroy semantics.

4. Scalability

Enterprise tables and graphs contain millions of rows or nodes. SSL training becomes computationally expensive.

5. Evaluation Metrics

Representation quality is usually measured via downstream tasks. Direct SSL evaluation benchmarks are limited.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

EMERGING RESEARCH DIRECTIONS

Self-supervised structured learning is rapidly evolving.

1. Multimodal Structured SSL

Future systems combine structured tables with text, images, and graphs. Example: medical tables + clinical notes.

2. Foundation Models for Tabular Data

Large pre-trained models for tables may learn universal structured representations similar to language models.

3. Causal Representation Learning

SSL can incorporate causal relations among attributes rather than correlations, improving robustness.

4. Automated Pretext Design

Meta-learning may automatically generate optimal SSL tasks for specific structured datasets.

5. Privacy-Preserving Structured SSL

Federated and encrypted SSL allows learning from distributed structured data without sharing raw records.

COMPARATIVE ANALYSIS OF STRUCTURED SSL APPROACHES

Data Type	Main SSL Method	Strength	Limitation
Tabular	Masked modeling	Feature relations	Ignores row similarity
Tabular	Contrastive	Robust embeddings	Needs augmentation
Graph	Contrastive GNN	Structural awareness	Heavy computation
Graph	Autoencoder	Link prediction	Reconstruction bias
Temporal	Forecast SSL	Temporal capture	Horizon sensitivity
Relational	Schema SSL	Multi-table relations	Complex modeling

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

FUTURE OUTLOOK

Self-supervised representation learning for structured data is expected to become central in data-centric AI. With increasing availability of unlabeled enterprise and scientific datasets, SSL provides scalable learning without annotation cost. Integration with graph neural networks, transformers, and relational modeling will likely create unified structured learning architectures.

Large tabular foundation models may emerge that generalize across domains such as finance, healthcare, and engineering. Moreover, combining SSL with causal inference and privacy preservation will enhance trustworthiness in critical applications.

CONCLUSION

Structured data dominates real-world decision systems, yet representation learning from such data remains challenging due to heterogeneity, relational dependencies, and limited labels. Self-supervised learning offers a powerful paradigm to learn meaningful representations by exploiting intrinsic structure of tables, graphs, and temporal data. This paper reviewed major SSL frameworks including masked modeling, contrastive learning, and generative approaches across tabular, graph, relational, and temporal structured datasets. Comparative analysis and application domains such as anomaly detection, recommender systems, healthcare analytics, and predictive maintenance were discussed. Key challenges include augmentation design, scalability, missing data handling, and evaluation metrics. Emerging directions such as multimodal structured SSL, foundation models, causal representation learning, and privacy-preserving methods indicate strong future potential. Self-supervised representation learning is therefore poised to become a core methodology for intelligent structured data analytics.

REFERENCES

1. Bengio, Y., Courville, A., & Vincent, P. Representation learning: A review and new perspectives. IEEE TPAMI, 2013.
2. Le-Khac, P., Healy, G., & Smeaton, A. Contrastive representation learning: A framework and review. IEEE Access, 2020.
3. Zhou, J., Cui, G., Zhang, Z., et al. Graph neural networks: A review of methods and applications. AI Open, 2020.
4. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. A simple framework for contrastive

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved

learning. ICML, 2020.

5. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. Revisiting deep learning models for tabular data. NeurIPS, 2021.
6. Huang, X., Zhang, J., Li, D., & Li, P. TabTransformer: Tabular data modeling using contextual embeddings. NeurIPS, 2020.
7. You, Y., Chen, T., Sui, Y., et al. Graph contrastive learning with augmentations. NeurIPS, 2020.
8. Vashishth, S., Sanyal, S., Nitin, V., & Talukdar, P. Composition-based multi-relational graph embeddings. ICLR, 2020.
9. Zerveas, G., Jayaraman, S., Patel, D., et al. Transformer models for multivariate time series. KDD, 2021.
10. Kossen, J., Band, N., Weiler, M., et al. Self-supervised learning for tabular data: A survey. 2021.

Page 73-89 © MANTECH PUBLICATIONS 2020. All Rights Reserved