DETAILED DESCRIPTION

[0014] FIG. 1 is a block diagram illustrating a machine learning environment 100 including a machine learning server 110 according to one embodiment. The environment 100 further includes multiple data sources 120 connected to the machine learning server 110 via a network 130. Although the illustrated environment 100 contains only one machine learning server 110 coupled to multiple data sources 120, embodiments can have multiple machine learning servers and a singular data source.

[0015] The data sources 120 provide electronic data to the machine learning server 110. A data source 120 may be a storage device such as a hard disk drive (HDD) or solid-state drive (SSD), a computer managing and providing access to multiple storage devices, a storage area network (SAN), a database, or a cloud storage system. A data source 120 may also be a computer system that can retrieve data from another source. The data sources 120 may be remote from the machine learning server 110 and provide the data via the network 130. In addition, some or all data sources 120 may be directly coupled to the data analytics system and provide the data without passing the data through the network 130.

[0016] The data provided by the data sources 120 may be organized into data records (e.g., rows). Each data record includes one or more values. For example, a data record provided by a data source 120 may include a series of comma-separated values. The data describe information of relevance to an enterprise using the machine learning server 110. For example, data from a data source 120 can describe computer-based interactions (e.g., click tracking data) with content accessible on websites and/or with applications. As another example, data from a data source 120 can describe customer transactions online and/or in stores. The enterprise can be in one or more of various industries, such as manufacturing, sales, financing, and banking.

[0017] The machine learning server 110 is a computer-based system utilized for constructing machine learning models and providing the machine learning models to make predictions based on data. Example predictions include whether or not a customer will make a transaction within a time period, whether or not a transaction is fraudulent, whether or not a user will perform a computer-based interaction, etc. The data are collected, gathered, or otherwise accessed from one or more of the multiple data sources 120 via the network 130. The machine learning server 110 can implement scalable software tools and hardware resources employed in accessing, preparing, blending, and analyzing data from a wide variety of data sources 120. The machine learning server 110 can be a computing device used to

CLAIMS

1. A computer implemented method, comprising:

receiving a dataset from a data source;

selecting primitives from a pool of primitives based on the received dataset, each of the selected primitives configured to be applied to at least a portion of the dataset to synthesize one or more features, wherein selecting the primitives from the pool of primitives comprises:

generating a semantic representation of the received dataset; and

selecting primitives associated with attributes matching the semantic representation of the received dataset;

synthesizing a plurality of features by applying the selected primitives to the received dataset;

iteratively evaluating the plurality of features to remove some features from the plurality of features to obtain a subset of features, each iteration comprising:

evaluating usefulness of at least some features of the plurality of features by applying a different portion of the dataset to the evaluated features; and

removing some of the evaluated features based on the usefulness of the evaluated features to produce the subset of features;

determining an importance factor for each feature of the subset of features; and

generating a machine learning model based on the subset of features and the importance factor of each feature of the subset of features, the machine learning model configured to be used to make a prediction based on new data.

2. The method of claim 1, wherein selecting the primitives from the plurality of primitives based on the received dataset further comprises:

generating a representative vector from the received dataset; and

inputting the representative vector into a machine learned model, the machine learned model outputting the selected primitives based on the representative vector.

3. The method of claim 1, wherein iteratively evaluating the plurality of features to remove some features from the plurality of features to obtain the subset of features comprises:

applying the plurality of features to a first portion of the dataset to determine a first usefulness score of each of the plurality of features;

removing some of the plurality of features based on the first usefulness score of each of the plurality of features to obtain a preliminary subset of features;

applying the preliminary subset of features to a second portion of the dataset to determine a second usefulness score of each of the preliminary subset of features; and

removing some of the preliminary subset of features from the preliminary subset of features based on the second usefulness score of each of the preliminary subset of features.

4. The method of claim 1, wherein determining the importance factor for each of the subset of features comprises:

ranking the subset of features by inputting the subset of features and a first portion of the dataset into a machine learning model, the machine learning model outputting a first ranking score for each of the subset of features; and

determining the importance factors of the subset of features based on their ranking scores.

5. The method of claim 4, further comprising:

ranking the subset of features by inputting the subset of features and a second portion of the dataset into a machine learning model, the machine learning model outputting a second ranking score for each of the subset of features;

determining a second importance factor for each of the subset of features based on the ranking score of the feature; and

adjusting the importance score for each of the subset of features based on the second importance score of the feature.

6. The method of claim 1, wherein synthesizing the plurality of features based on the subset of primitives and the received dataset comprises:

for each primitive in the subset:

identifying one or more variables in the dataset; and

applying the primitive to the one or more variables to generate one or more features of the plurality of features.

7. The method of claim 1, further comprising:

displaying a user interface that includes a feature selection tool configured to receive user input defining at least one of:

a maximum feature depth;

a maximum number of features in the subset of features; or

a temporal constraint for the subset of features; and

generating the subset of features based on the user input to the feature selection tool.

8. A system, comprising:

a computer processor for executing computer program instructions; and

a non-transitory computer-readable memory storing computer program instructions executable by the computer processor to perform operations comprising:

receiving a dataset from a data source;

selecting primitives from a pool of primitives based on the received dataset, each of the selected primitives configured to be applied to at least a portion of the dataset to synthesize one or more features, wherein selecting the primitives from the pool of primitives comprises:

generating a semantic representation of the received dataset; and

selecting primitives associated with attributes matching the semantic representation of the received dataset;

synthesizing a plurality of features by applying the selected primitives to the received dataset;

iteratively evaluating the plurality of features to remove some features from the plurality of features to obtain a subset of features, each iteration comprising:

evaluating usefulness of at least some features of the plurality of features by applying a different portion of the dataset to the evaluated features; and

removing some of the evaluated features based on the usefulness of the evaluated features to produce the subset of features;

determining an importance factor for each feature of the subset of features; and

generating a machine learning model based on the subset of features and the importance factor of each feature of the subset of features, the machine learning model configured to be used to make a prediction based on new data.

9. The system of claim 8, wherein selecting the subset of primitives from the plurality of primitives based on the received dataset further comprises:

generating a representative vector from the received dataset; and

inputting the representative vector into a machine learned model, the machine learned model outputting the selected primitives based on the representative vector.

10. The system of claim 8, wherein iteratively evaluating the plurality of features to remove some features from the plurality of features to obtain the subset of features comprises:

applying the plurality of features to a first portion of the dataset to determine a first usefulness score of each of the plurality of features;

removing some of the plurality of features based on the first usefulness score of each of the plurality of features to obtain a preliminary subset of features;

applying the preliminary subset of features to a second portion of the dataset to determine a second usefulness score of each of the preliminary subset of features; and

removing some of the preliminary subset of features from the preliminary subset of features based on the second usefulness score of each of the preliminary subset of features.

11. The system of claim 8, wherein determining the importance factor for each of the subset of features comprises:

ranking the subset of features by inputting the subset of features and a first portion of the dataset into a machine learning model, the machine learning model outputting a first ranking score for each of the subset of features; and

determining the importance factors of the subset of features based on their ranking scores.

12. The system of claim 11, the operations further comprising:

ranking the subset of features by inputting the subset of features and a second portion of the dataset into a machine learning model, the machine learning model outputting a second ranking score for each of the subset of features;

determining a second importance factor for each of the subset of features based on the ranking score of the feature; and

adjusting the importance score for each of the subset of features based on the second importance score of the feature.

13. The system of claim 8, wherein synthesizing the plurality of features based on the subset of primitives and the received dataset comprises:

for each primitive in the subset:

identifying one or more variables in the dataset; and

applying the primitive to the one or more variables to generate one or more features of the plurality of features.

14. A non-transitory computer-readable memory storing executable computer program instructions for processing data blocks in a data analytics system, the instructions executable to perform operations comprising:

receiving a dataset from a data source;

selecting primitives from a pool of primitives based on the received dataset, each of the selected primitives configured to be applied to at least a portion of the dataset to synthesize one or more features, wherein selecting the primitives from the pool of primitives comprises:

generating a semantic representation of the received dataset; and

selecting primitives associated with attributes matching the semantic representation of the received dataset;

synthesizing a plurality of features by applying the selected primitives to the received dataset;

iteratively evaluating the plurality of features to remove some features from the plurality of features to obtain a subset of features, each iteration comprising:

evaluating usefulness of at least some features of the plurality of features by applying a different portion of the dataset to the evaluated features; and

removing some of the evaluated features based on the usefulness of the evaluated features to produce the subset of features;

determining an importance factor for each feature of the subset of features; and

generating a machine learning model based on the subset of features and the importance factor of each feature of the subset of features, the machine learning model configured to be used to make a prediction based on new data.

15. The non-transitory computer-readable memory of claim 14, wherein selecting the primitives from the plurality of primitives based on the received dataset further comprises:

generating a representative vector from the received dataset; and

inputting the representative vector into a machine learned model, the machine learned model outputting the selected primitives based on the representative vector.

16. The non-transitory computer-readable memory of claim 14, wherein iteratively evaluating the plurality of features to remove some features from the plurality of features to obtain the subset of features comprises:

applying the plurality of features to a first portion of the dataset to determine a first usefulness score of each of the plurality of features;

removing some of the plurality of features based on the first usefulness score of each of the plurality of features to obtain a preliminary subset of features;

applying the preliminary subset of features to a second portion of the dataset to determine a second usefulness score of each of the preliminary subset of features; and

removing some of the preliminary subset of features from the preliminary subset of features based on the second usefulness score of each of the preliminary subset of features.

17. The non-transitory computer-readable memory of claim 14, wherein determining the importance factor for each of the subset of features comprises:

ranking the subset of features by inputting the subset of features and a first portion of the dataset into a machine learning model, the machine learning model outputting a first ranking score for each of the subset of features; and

determining the importance factors of the subset of features based on their ranking scores.

18. The non-transitory computer-readable memory of claim 17, the operations further comprising:

ranking the subset of features by inputting the subset of features and a second portion of the dataset into a machine learning model, the machine learning model outputting a second ranking score for each of the subset of features;

determining a second importance factor for each of the subset of features based on the ranking score of the feature; and

adjusting the importance score for each of the subset of features based on the second importance score of the feature.

19. The non-transitory computer-readable memory of claim 14, wherein synthesizing the plurality of features based on the subset of primitives and the received dataset comprises:

for each primitive in the subset;

identifying one or more variables in the dataset; and

applying the primitive to the one or more variables to generate one or more features of the plurality of features.

20. The non-transitory computer-readable memory of claim 14, the operations further comprising:

displaying a user interface that includes a feature selection tool configured to receive user input defining at least one of:

a maximum feature depth;

a maximum number of features in the subset of features; or

a temporal constraint for the subset of features; and

generating the subset of features based on the user input to the feature selection tool.