

WHAT IS CLAIMED IS:

1. In a system for preparing a plurality of historical data records for use in training a primary model machine learning algorithm, wherein each historical data record includes a plurality of fields containing data values designated as inputs for training the primary model machine learning algorithm to generate an output of interest, a computer-implemented method of preparing the designated inputs in the plurality of historical data records in order to increase the utility and accuracy of the trained primary model machine learning algorithm when the designated inputs in the historical data records include invalid data values, the method comprising:

segregating a base dataset containing the historical data records into a faulty dataset having incomplete data records with invalid data values and a clean dataset having complete data records with no invalid data values;

storing the faulty dataset and the clean dataset in a computer memory;

producing from the stored clean dataset at least one of (i) plural computer-implemented secondary model machine learning algorithms trained to generate from values of the respective fields designated as inputs in plural complete data records a replacement value for a single invalid data value in a corresponding field designated as an input in an incomplete data record, and (ii) a computer-implemented tertiary model machine learning clustering algorithm trained to generate from plural complete data records comprising all values of fields designated as inputs replacement values for multiple invalid data values designated as inputs in an incomplete data record; and

using a computer-implemented program to create augmented training data records by substituting the replacement data values for at least some of the respective invalid data values designated as inputs in the stored faulty dataset, whereby said augmented training data records can be used with complete data records in the clean dataset to train the primary model machine learning algorithm to improve the accuracy thereof when generating an output of interest from a new data record.

2. A system as in claim 1, wherein the method further comprises:

training the primary model machine learning algorithm using the augmented training data records;

obtaining a new data record with fields corresponding to respective ones of the fields in the historical data records designated as inputs;

storing the new data record in a computer memory;

completing the new data record by applying to the stored new data record the secondary model machine learning algorithm to generate a replacement value for data in the new data record with a single field containing an invalid data value, and the tertiary model machine learning clustering algorithm to generate replacement values for data in the new data record with multiple fields containing invalid data values; and

using the trained primary model machine learning algorithm to generate the output of interest from the new data record.

3. A system as in claim 1, wherein at least one field in an historical data record is designated as an output of interest comprising a target data value in numeric form, and the primary machine learning algorithm uses supervised learning to fit a curve relating the data values of fields designated as inputs in the historical data records to the numeric target data values in the historical data records.

4. A system as in claim 1, wherein at least one field in an historical data record is designated as an output of interest comprising target data values in the form of two or more discrete classes, and the primary machine learning algorithm uses supervised learning to maximize the probability that the values of the data designated as inputs in the historical data records determine that the data record is a member of one of the two or more discrete classes comprising the target data values in the historical data records.

5. A system as in claim 1, wherein the primary machine learning algorithm output of interest is an identification of a collection of data records whose values designated as inputs are more similar to other data records in the collection than said input values are similar to values designated as inputs in data records which are not in the collection.

6. A system as in claim 1, wherein the secondary model machine learning algorithm comprises at least one of a prediction model for generating replacement values for fields having values in continuous numeric form and a classification model for generating replacement values for fields having values in the form of discrete classes.

7. A system as in claim 1, wherein the secondary model machine learning algorithm comprises a multi-layer feed-forward neural network trained by back-propagation.

8. A system as in claim 1, wherein the tertiary model machine learning algorithm comprises a self-organizing map characterized by a plurality of clusters based on the total number of historical data records.

9. A system as in claim 1, wherein the method of preparing the designated inputs in the plurality of historical data records for training of the primary model further comprises:

using a heuristic analysis to identify any of the fields designated as inputs for the primary model machine learning algorithm that contain data values having no utility for training the primary machine learning algorithm to generate the output of interest;

creating a reduced clean dataset without the fields containing data values identified as having no utility for training the primary machine learning algorithm to generate the output of interest and storing the reduced clean dataset in a computer memory;

creating an auxiliary clean dataset without any fields representing the output of interest and storing the auxiliary clean dataset; and

using the stored auxiliary clean dataset as training data for the plural secondary model machine learning algorithms and the tertiary model machine learning clustering algorithm.

10. A method of using a computer-implemented primary model machine learning algorithm trained with a plurality of historical data records, wherein each historical data record includes a plurality of fields designated as inputs for training the primary model machine learning algorithm to generate an output of interest, wherein the method generates a corresponding output of interest from a new data record with a plurality of fields corresponding to the fields in the historical data records designated as inputs when one or more of the fields in the new data record contains an invalid data value, the method comprising:

using one of a computer-implemented secondary model machine learning algorithm trained using the historical data records to generate a replacement value for a single field containing an invalid data value, and a computer-implemented tertiary model machine learning clustering algorithm trained using the historical data records to generate replacement values for a data record with multiple fields containing invalid data values;

completing the new data record by substituting the one or more replacement values corresponding to the data values in respective fields of the new data record containing an invalid data value; and

using the primary model machine learning algorithm to generate from the completed new data record the output of interest associated with the new data record.

11. A method as in claim 10, further comprising:

accessing a base dataset with a plurality of the historical data records;

segregating from the stored base dataset a clean dataset having complete historical data records with no invalid data values;

storing the clean dataset in a computer memory;

producing from the stored clean dataset the secondary model machine learning algorithm and the tertiary model machine learning clustering algorithm.

12. A method as in claim 11, further comprising;

storing in a computer memory a faulty dataset having incomplete historical data records with invalid data values;

using a computer-implemented program to create augmented training data records by substituting the replacement data values for at least some of the respective invalid data values in data records in the stored faulty dataset and storing the augmented training data records in a computer memory; and

training the primary model machine learning algorithm using the data records in the clean dataset combined with the augmented training data records.

13. A method of producing plural secondary model machine learning algorithms for replacing invalid data values in a stored base dataset that includes a plurality of historical data records each having a plurality of fields designated as inputs for training a primary model machine learning algorithm to generate an output of interest, the method comprising:

using a heuristic analysis to identify any of the fields in the historical data records designated as inputs for the primary model machine learning algorithm that contain data values having no utility for training the primary machine learning algorithm to generate the output of interest;

storing in a computer memory a field status data structure with an entry associated with each field in data records in the base dataset, wherein each field status entry includes a field number element indicating the position of the field in data records in the stored base dataset and an elimination code element indicating whether the data values contained in said field have utility for training the primary model to generate an output of interest;

creating a clean dataset by eliminating from the base dataset any data records with invalid data values;

creating a reduced clean dataset by eliminating from data records in the clean dataset any field having an elimination code element associated therewith that indicates the field has no utility for training the primary model;

creating an auxiliary clean dataset by eliminating from data records in the reduced clean dataset any field with a data value designated as the output of interest of the primary model machine learning algorithm; and

using the data records in the auxiliary clean dataset as training data for plural computer-implemented secondary model machine learning algorithms trained by sequentially designating all but one of the fields of said data records as containing training input values and designating said one remaining field data value as the training output value for the secondary model machine learning algorithm for that field;

storing in a computer memory a replacement model data structure with an entry associated with each field in the base dataset, the position of each entry in the replacement model data structure being indicative of the field number element in the corresponding field status data structure, and each entry including (a) a field index element indicating the position for the field value in a data record organized for use by a primary model machine learning algorithm, and (b) a secondary model element indicating either (i) the trained secondary model machine learning algorithm associated with said field, or (ii) that no secondary model machine learning algorithm is associated with said field.

14. A method as in claim 13, wherein the method further comprises including in each field status data structure entry a type code element indicating (i) whether or not the data values in the field associated with said entry will be used to produce a secondary model machine learning algorithm, and (ii) whether the data values of the training outputs of data records used to produce

the secondary model machine learning algorithm for said field are in continuous numeric form or discrete class form.

15. A method as in claim 14, wherein the secondary model machine learning algorithms comprise multi-layer feed-forward neural networks trained by back-propagation.

16. A method for training a primary machine learning algorithm using trained secondary model machine learning algorithms produced as in claim 13, the method comprising:

- (a) creating a faulty dataset comprising faulty historical data records eliminated from the base dataset and storing the faulty dataset in a computer memory;
- (b) identifying in the faulty dataset a faulty data record with a single field having an invalid data value;
- (c) accessing the entry in the replacement model data structure corresponding to the single field having an invalid data value to determine whether or not said single field in said faulty data record has a trained secondary model machine learning algorithm associated with said field;
- (d) using said trained secondary model machine learning algorithm associated with said field accessed in step (c) to create a replacement data value;
- (e) substituting the replacement value for the invalid data value in the faulty data record to create an augmented training data record;
- (f) repeating steps (b) through (e) a plurality of times for different data records in the faulty dataset; and
- (g) using the data records in the clean dataset combined with the augmented training data records to train the primary model machine learning algorithm.

17. A method for using the primary machine learning algorithm trained in accordance with claim 16 to generate an output of interest from a new data record with input fields corresponding to the input fields in the base dataset, the method comprising:

- (a) identifying in the new data record a single field having an invalid data value;

(b) accessing the entry in the replacement model data structure corresponding to the single field having an invalid data value to determine whether or not said single field in said faulty data record has a trained secondary model machine learning algorithm associated with said field;

(c) using said trained secondary model machine learning algorithm associated with said field accessed in step (b) to create a replacement data value for the invalid data value in the new data record;

(d) substituting the replacement data value for the invalid data value to create an augmented new data record; and

(e) using the primary model machine learning algorithm to generate the output of interest from the augmented new data record.

18. A method of producing a tertiary model machine learning algorithm for replacing invalid data values in a stored base dataset that includes a plurality of historical data records each having a plurality of fields designated as inputs for training a primary model machine learning algorithm to generate an output of interest, the method comprising:

using a heuristic analysis to identify any of the fields in the historical data records designated as inputs for the primary model machine learning algorithm that contain data values having no utility for training the primary machine learning algorithm to generate the output of interest;

storing in a computer memory a field status data structure with an entry associated with each field in the base dataset, wherein each field status entry includes a field number element indicating the position of the field in data records in the stored base dataset and an elimination code element indicating whether the data values contained in said field have utility for training the primary model to generate an output of interest;

creating a clean dataset by eliminating from the base dataset any data records with invalid data values;

creating a reduced clean dataset by eliminating from data records in the clean dataset any field having associated therewith an elimination code element that indicates the field has no utility for training the primary model;

creating an auxiliary clean dataset by eliminating from data records in the reduced clean dataset any field with a data value designated as the output of interest of the primary model machine learning algorithm; and

using the data records in the auxiliary clean dataset as training data for a computer-implemented tertiary model machine learning clustering algorithm trained by using all of the data values in data records in the auxiliary clean dataset as training input values; and

storing in a computer memory a replacement model data structure with a tertiary model entry identifying the trained tertiary model machine learning clustering algorithm.

19. A method for training a primary machine learning algorithm using the trained tertiary model machine learning algorithm produced as in claim 18, the method comprising:

(a) creating a faulty dataset comprising faulty historical data records eliminated from the base dataset and storing the faulty dataset in a computer memory;

(b) identifying in the faulty dataset a faulty data record with multiple fields having invalid data values;

(c) accessing the tertiary model entry in the replacement model data structure;

(d) using said trained tertiary model machine learning algorithm accessed in step (c) to create replacement data values for the invalid data values in the faulty data record;

(e) substituting the replacement values for the invalid data values in the faulty data record to create an augmented training data record;

(f) repeating steps (b) through (e) a plurality of times for different data records in the faulty dataset; and

(g) using the data records in the clean dataset combined with the augmented training data records to train the primary model machine learning algorithm.

20. A method as in claim 19, wherein the trained tertiary model machine learning clustering algorithm comprises a self-organizing map with a plurality of nodes having respective weights corresponding to data values in the auxiliary clean dataset, and step (d) comprises:

calculating a similarity metric defining a distance feature between the weights of each node of the self-organizing map and each of the fields with valid data values in the faulty data record;

using the distance feature to designate a predetermined number of winning nodes having weights closest to the corresponding values of respective fields in the faulty data record with invalid data values;

calculating an average of the weights of the winning nodes associated with each of the invalid data values; and

using the average weights to generate respective replacement values for each of the invalid data values in the fields associated with the weights.

21. A method for using the primary machine learning algorithm trained in accordance with claim 19 to generate an output of interest from a new data record with input fields corresponding to the input fields in the base dataset, the method comprising:

identifying in the new data record multiple fields having invalid data values;

using the trained tertiary model machine learning algorithm to create replacement data values for the invalid data values in the new data record;

substituting the replacement data values for the invalid data values to create an augmented new data record; and

using the primary model machine learning algorithm to generate the output of interest from the augmented new data record.

22. A method of producing plural secondary model machine learning algorithms and a tertiary model machine learning algorithm for replacing invalid data values in a stored base dataset that includes a plurality of historical data records each having a plurality of fields designated as

inputs for training a primary model machine learning algorithm to generate an output of interest, the method comprising:

using a heuristic analysis to identify any of the fields in the historical data records designated as inputs for the primary model machine learning algorithm that contain data values having no utility for training the primary machine learning algorithm to generate the output of interest;

storing in a computer memory a field status data structure with an entry associated with each field in data records in the base dataset, wherein each field status entry includes a field number element indicating the position of the field in data records in the stored base dataset and an elimination code element indicating whether the data values contained in said field have utility for training the primary model to generate an output of interest;

creating a clean dataset by eliminating from the base dataset any data records with invalid data values;

creating a reduced clean dataset by eliminating from data records in the clean dataset any field having an elimination code element associated therewith that indicates the field has no utility for training the primary model;

creating an auxiliary clean dataset by eliminating from data records in the reduced clean dataset any field with a data value designated as the output of interest of the primary model machine learning algorithm;

using the data records in the auxiliary clean dataset as training data for plural computer-implemented secondary model machine learning algorithms trained by sequentially designating all but one of the fields of said data records as containing training input values and designating said one remaining field data value as the training output value for the secondary model machine learning algorithm for that field;

using the data records in the auxiliary clean dataset as training data for a computer-implemented tertiary model machine learning clustering algorithm trained by using all of the data values in data records in the auxiliary clean dataset as training input values; and

storing in a computer memory a replacement model data structure with a tertiary model entry identifying the trained tertiary model machine learning algorithm and plural secondary model entries, each secondary model being associated with a respective field in the base dataset, wherein the position of each entry in the replacement model data structure is indicative of the field number element in the corresponding field status data structure, and each entry includes (a) a field index element indicating the position for the field value in a data record organized for use by a primary model machine learning algorithm, and (b) a secondary model element indicating either (i) the trained secondary model machine learning algorithm associated with said field, or (ii) that no secondary model machine learning algorithm is associated with said field.

23. A method as in claim 22, wherein the method further comprises including in each field status data structure entry a type code element indicating (i) whether or not the data values in the field associated with said entry will be used to produce a secondary model machine learning algorithm, and (ii) whether the data values of the training outputs of data records used to produce the secondary model machine learning algorithm for said field are in continuous numeric form or discrete class form.

24. A method as in claim 23, wherein the secondary model machine learning algorithms comprise multi-layer feed-forward neural networks trained by back-propagation.

25. A method for training a primary machine learning algorithm using trained secondary model machine learning algorithms and the trained tertiary model machine learning algorithm produced as in claim 22, the method comprising:

(a) creating a faulty dataset comprising faulty historical data records eliminated from the base dataset and storing the faulty dataset in a computer memory;

(b) identifying in the faulty dataset a faulty data record with a single field having an invalid data value;

(c) accessing the entry in the replacement model data structure corresponding to the single field having an invalid data value to determine whether or not said single field in said faulty data record has a trained secondary model machine learning algorithm associated with said field;

(d) using said trained secondary model machine learning algorithm associated with said field accessed in step (c) to create a replacement data value;

(e) substituting the replacement value for the invalid data value in the faulty data record to create an augmented training data record;

(f) repeating steps (b) through (e) a plurality of times for different data records in the faulty dataset;

(g) identifying in the faulty dataset a faulty data record with multiple fields having invalid data values;

(h) accessing the tertiary model entry in the replacement model data structure;

(i) using said trained tertiary model machine learning algorithm accessed in step (h) to create replacement data values for the invalid data values in the faulty data record;

(j) substituting the replacement values for the multiple invalid data values in the faulty data record to create an augmented training data record; and

(k) using the data records in the clean dataset combined with the augmented training data records created in steps (e) and (j) to train the primary model machine learning algorithm.

26. A method as in claim 25, wherein the trained tertiary model machine learning clustering algorithm comprises a self-organizing map with a plurality of nodes having respective weights corresponding to data values in the auxiliary clean dataset, and step (i) comprises:

calculating a similarity metric defining a distance feature between the weights of each node of the self-organizing map and each of the fields with valid data values in the faulty data record;

using the distance feature to designate a predetermined number of winning nodes having weights closest to the corresponding values of respective fields in the faulty data record with invalid data values;

calculating an average of the weights of the winning nodes associated with each of the invalid data values; and

using the average weights to generate respective replacement values for each of the invalid data values in the fields associated with the weights.