Claims

1. A feature creation system that facilitates data mining operations used to produce training data that can be used to improve the accuracy of machine learning models, the system comprising:

one or more hardware processors configured to execute software instructions; one or more storage devices storing software instructions configured to execution by the one or more hardware computer processors in order to cause the feature creation system to;

store in a storage unit a first table and a second table, the first table including multiple rows, having a value for a prediction object and a value for a first attribute, and the second table including multiple rows having a value for a second attribute, the first and second attributes used as features for training a machine learning model to generate prediction results for the prediction object;

determine a similarity between one or more of the values for the first and second attributes based on a similarity-function used to calculate a degree of similarity between the one or more values for the first attribute and the second attribute, and a similarity condition that includes a threshold value for the degree of similarity;

generate from the first table and the second table, feature candidates that influence the prediction object based on a joining condition that includes the degree of similarity between the one or more values for the first and second attributes and a reduction condition, the joining condition used to join one or more records included in the first table and one or more records included in the second table, the reduction condition used to determine a reduction method for reducing a plurality of records and columns in the second table, the feature candidates include values for the first and second attributes that satisfy the similarity condition;

select an optimum feature for prediction from the feature candidates; and

train a prediction model using the optimum feature.

2. The feature creation system of claim 1, wherein the joining condition is selected from a plurality of joining conditions and the reduction condition is selected from a plurality of reduction conditions.

3. The feature creation system of claim 2, wherein the software instructions are further configured to generate the feature candidates using the plurality of the joining conditions and the plurality of the reduction conditions, and the number of the feature candidates is equivalent to a number of combinations consisting of at least one

joining condition selected from the plurality of joining conditions and at least one reduction condition selected from the plurality of reduction conditions.

4. The feature creation system of claim 1, wherein the first attribute and the second attribute are geographical attributes, and the similarity function provides a high value for the similarity based on a near distance between the geographical attributes.

5. The feature creation system of claim 1, wherein the first attribute and the second attribute are temporal attributes, and the similarity function provides a high value for the similarity based on a small time difference between the temporal attributes.

6. The feature creation system of claim 1, wherein the first attribute and the second attribute are character string attributes, and the similarity function provides a high value for the similarity based on a high degree of matching of texts represented by the character string attributes.

7. The feature creation system of claim 1, wherein the first attribute and the second attribute are structural attributes, and the function provides calculates a high value for the similarity based on a near distance between the structural attributes an a common node.

8. The feature creation system of claim 1, wherein the software instructions are further configured to generate:

a feature descriptor using the joining condition, the reduction condition and a filtering condition.

9. The feature creation system of claim 8, wherein the filtering condition is used to filters records using a condition from the second table.

10. A method of feature creation that facilitates data mining operations used to produce training data that can be used to improve the accuracy of machine learning models, the method comprising:

storing in a storage unit a first table and a second table, the first table including multiple rows, having a value for a prediction object and a value for a first attribute, and the second table including multiple rows having a value for a second attribute, the first and second attributes used as features for training a machine

learning model to generate prediction results for the prediction object;

determining a similarity between one or more of the values for the first and second attributes based on a similarity function used to calculate a degree of similarity between the one or more values for the first attribute and the second attribute, and similarity condition that includes a threshold value for the degree of similarity;

generating from the first table and the second table, feature candidates that influence the prediction object based on a joining condition that includes the degree of similarity between the one or more values for the first and second attributes and a reduction condition, the joining condition used to join one or more records included ni the first table and one or more records included in the second table, the reduction condition used to determine a reduction method for reducing a plurality of records and columns in the second table, the feature candidates include values for the first and second attributes that satisfy the similarity condition; and

training a prediction model using the optimum feature.

11. The method of claim 10, wherein the joining condition is selected from a plurality of joining conditions and the reduction condition is selected from a plurality of reduction conditions.

12. The method of claim 11, wherein the feature candidates are generated using the plurality of the joining conditions and the plurality of the reduction conditions, wherein the number of the feature candidates generated is equivalent to a number of combinations consisting of at least one joining condition selected from the plurality of joining conditions and at least one reduction condition selected from the plurality of reduction conditions.

13. The method of claim 10, wherein the first attribute and the second attribute are geographical attributes, and the similarity function calculates a high value for the similarity based on a near distance between the geographical attributes.

14. The method of claim 10, wherein the first attribute and the second attribute are temporal attributes, and the similarity function calculates a high value for the similarity based on a small time difference between the temporal attributes.

15. The method of claim 10, wherein the first attribute and the second attribute are character string attributes, and the similarity function calculates a high value for the similarity based on a high degree of matching of texts represented by the

character string attributes.

16. The method of claim 10, wherein the first attribute and the second attribute are structural attributes, and the similarity function calculates a high value for the similarity based on a near distance between the structural attributes an a common node.

17. The method of claim 10 further comprising:

creating a feature descriptor based on the joining condition, the reduction condition, and a filtering condition.

18. The method of claim 17, wherein the filtering condition filters records using a condition from the second table.