



## (51) International Patent Classification:

G06N 3/045 (2023.01) G06N 3/091 (2023.01)

G06N 3/0475 (2023.01) G05D 1/00 (2006.01)

G06N 3/0895 (2023.01) G06N 5/022 (2023.01)

## (21) International Application Number:

PCT/EP2022/076440

## (22) International Filing Date:

22 September 2022 (22.09.2022)

## (25) Filing Language:

English

## (26) Publication Language:

English

## (30) Priority Data:

22170719.3 29 April 2022 (29.04.2022) EP

## (71) Applicant: NEC LABORATORIES EUROPE GMBH

[DE/DE]; Kurfürsten-Anlage 36, 69115 Heidelberg (DE).

(72) Inventors: **FUERST, Jonathan**; NEC Laboratories Europe GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg (DE). **CHENG, Bin**; NEC Laboratories Europe GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg (DE).

(74) Agent: **ULLRICH & NAUMANN**; Schneidmühlstraße 21, 69115 Heidelberg (DE).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: A COMPUTER-IMPLEMENTED ACTIVE LEARNING METHOD IN THE FIELD OF AN IMBALANCED MATCHING SITUATION AND A CORRESPONDING SYSTEM

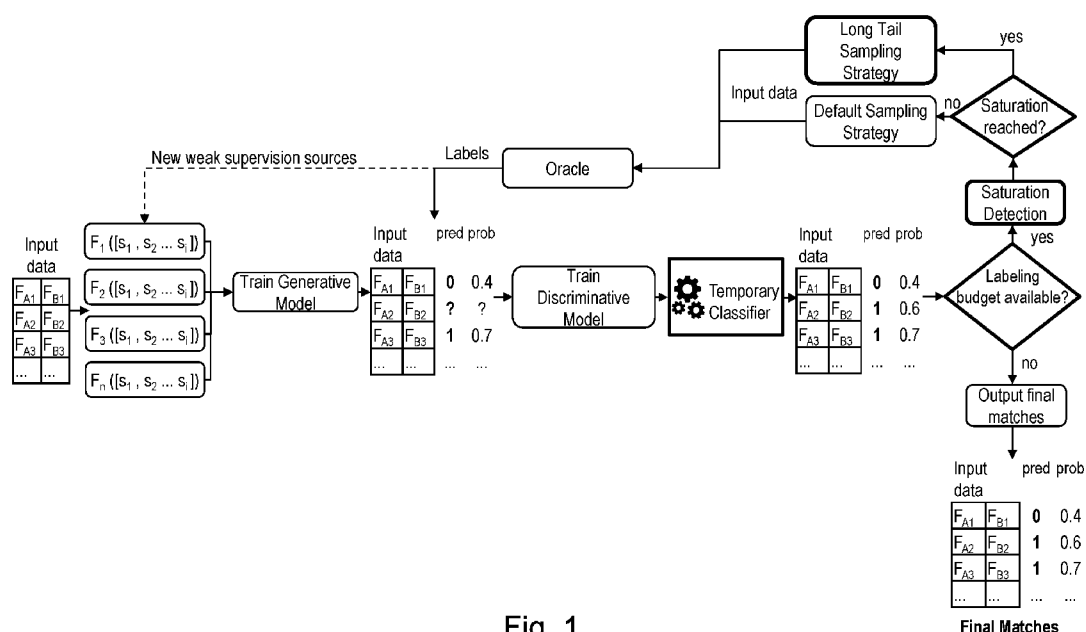


Fig. 1

(57) Abstract: For providing a high degree of performance of a computer-implemented active learning method by simple means a computer-implemented active learning method in the field of an imbalanced matching situation, for finding at least one relationship between data points in different data sets, is provided, comprising the following steps: executing an uncertainty-based sampling strategy on a defined set of data points; performing a detection of a defined saturation level in an improvement of a model performance, wherein the detection is based on outputs of a generative model and a discriminative model; further executing the uncertainty-based sampling strategy, if such a saturation level is not detected, and switching to executing a long-tail sampling strategy, if such a saturation level is detected; and outputting a final model for prediction, if a defined budget of labels for the data points is depleted. Further, a correspond-

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS,  
ZA, ZM, ZW.

**(84) Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

**Published:**

— *with international search report (Art. 21(3))*

ing system for carrying out the computer-implemented active learning method in the field of an imbalanced matching situation, for finding at least one relationship between data points in different data sets, is provided, comprising: executing means for executing an uncertainty-based sampling strategy on a defined set of data points; performing means for performing a detection of a defined saturation level in an improvement of a model performance, wherein the detection is based on outputs of a generative model and a discriminative model; executing means for further executing the uncertainty-based sampling strategy, if such a saturation level is not detected, and switching means for switching to executing means for executing a long-tail sampling strategy, if such a saturation level is detected; and outputting means for outputting a final model for prediction, if a defined budget of labels for the data points is depleted.

## A COMPUTER-IMPLEMENTED ACTIVE LEARNING METHOD IN THE FIELD OF AN IMBALANCED MATCHING SITUATION AND A CORRESPONDING SYSTEM

5      The present invention relates to a computer-implemented active learning method in  
the field of an imbalanced matching situation, for finding at least one relationship  
between data points in different data sets.

10      Further, the present invention relates to a corresponding system for carrying out the  
computer-implemented active learning method.

Corresponding prior art documents are listed as follows:

15      [1] Vlachos, Andreas. "A stopping criterion for active learning." Computer Speech &  
Language 22, no. 3, 2008: 295-312.

20      [2] Laws, F. and Schütze, H., 2008, August. Stopping criteria for active learning of  
named entity recognition. In Proceedings of the 22nd International Conference on  
Computational Linguistics, Coling, 2008, pp. 465-472.

[3] Broder, Andrei Z. "Identifying and filtering near-duplicate documents." In Annual  
Symposium on Combinatorial Pattern Matching, pp. 1-10. Springer, Berlin,  
Heidelberg, 2000.

25      [4] Rühling Cachay, Salva, Benedikt Boecking, and Artur Dubrawski. "End-to-End  
Weak Supervision." Advances in Neural Information Processing Systems 34, 2021.

30      [5] Stephen Mussmann, Robin Jia, and Percy Liang. 2020. On the Importance of  
Adaptive Data Collection for Extremely Imbalanced Pairwise Tasks. In Findings of  
the Association for Computational Linguistics: EMNLP 2020. 3400–3413.

Further prior art documents:

5 Sunita Sarawagi, Anuradha Bhamidipaty. "Interactive Deduplication using Active Learning", 2002. This document discloses an interactive deduplication method using an active learning approach and describes that the method includes the creation of weak supervision signals as noisy data input to a generative classifier model and the generative classifier model generates labels for the corresponding noisy data. The said labels and the output of the generative model are transferred to train the discriminative model to generate predictions. Furthermore, the uncertainty or saturation point is detected based on a comparison of the outputs of both the classifiers and then weighted sampling is performed. The final output of the last discriminative model is used for further predictions.

15 Rahul Vigneswaran, Marc T. Law, Vineeth N. Balasubramanian, Makarand Tapaswi. "Feature Generation for Long-tail Classification", November 10, 2021. This document discloses a feature generation method for long-tail classification using a deep learning approach and describes that the method includes the creation of input signals/samples as data input to a generative classifier model and the generative classifier model generates labels for the corresponding input data samples. The said labels and the output of the generative model are transferred to train the discriminative model to generate long-tail classification predictions. Furthermore, the document discloses a two-stage sampling approach including instance sampling and balanced sampling referred to as long-tail sampling.

25 Shyamgopal Karthik, Jerome Revaud, Chidlovskii Boris. "Learning From Long-Tailed Data With Noisy Labels", September 12, 2021. This document discloses a learning approach for long-tailed data with noisy labels and describes that the method includes the creation of input signals/samples as noisy data input to the first – generative classifier – model and the respective model generates labels for the corresponding input data samples. The said labels and the output of the first model are transferred to train the discriminative model to generate long-tail classification predictions. Furthermore, the document discloses a two-stage sampling approach including instance sampling and long-tail sampling.

In data integration/preparation, users are faced with several matching problems, from schema and ontology matching to entity matching/resolution. In essence, in these problems users are interested to find same-as/equal relationships between elements in disjoint datasets. E.g., users might need to integrate various data sources related to a non-residential building such as Industry Foundation Classes, IFC, building management system, BMS, data, energy management system, EMS, data, facility management system, FMS, Calendar entries, OpenStreetMap, Security System etc. to realize some novel application, such as advanced energy/carbon saving applications. This integration scenario involves matching different data schemas/ontologies and subsequently matching records that belong to the same real-world entity, e.g., records that refer to the same room in the building. To solve these matching problems, others have applied various solutions from heuristic and unsupervised, e.g., matching heuristics and lexical similarity based, approaches to recently increasingly supervised machine learning, ML, – most recent deep learning – based solutions, which tend to outperform previous works. However, one problem with supervised machine learning is the need for large, labeled datasets. One way to reduce this labeling bottleneck is Active Learning.

Active Learning describes ML methods which actively query an oracle, e.g., a system, expert user, database etc., to label data points. These labeled data points are then used to train a supervised ML model. As the oracle represents a bottleneck in terms of cost/effort, it is of paramount importance to select the “most informative” samples for labeling, opposed to just random sampling. The most informative samples are usually the ones which the trained model is most uncertain about, i.e., they are at the decision boundary. Different metrics have been proposed to quantify this uncertainty, e.g.:

- **Least Confidence Uncertainty**

Based on the difference between the most confident class prediction and 100% confidence, selecting the sample which has the lowest probability.

- **Smallest Margin Uncertainty/Margin of confidence**

Select based on the smallest difference/margin between the two topmost predictions.

- **Largest Margin Uncertainty**

Select based on the smallest difference/margin between the top and the bottom prediction.

- **Ratio of Uncertainty**

Based on the ratio of the top two predicted classes.

- **Entropy Reduction**

Entropy is the basis for mutual information, which quantifies relation between two things and the basis for relative entropy – kullback-leibler distance – and cross entropy. Higher values of entropy indicate greater uncertainty.

$$\text{Entropy} = -\sum p(x) \log(p(x))$$

Active learning can be bootstrapped with a random sampling strategy, which is not efficient for imbalanced datasets, such as the described matching problems, only < 0.1% of all combinations are matches, [5]. Thus, to bootstrap active learning faster, with less effort, weak supervision has been proposed [4]. Weak supervision provides an initial set of noisy weak outputs in this context. Active learning is then used to improve the quality of these noisy weak outputs by providing high-quality labels, i.e., highly accurate, for a small set of informative samples. However, the problem with using existing uncertainty measures in this scenario is that the query selection strategy is *biased to the provided weak-supervision signals*. This is especially an issue for imbalanced classification problems. For example, in entity matching, weak supervision signals, e.g., encoded in data programming-style labeling functions, might only use a subset of attributes of the entity in their matching heuristics, i.e., to decide if a combination is a match or a non-match.

If an un-used attribute is however a strong discriminator for the matching task, users might miss the affected entity pairs and consider them non-matches based on the other attributes. These false negatives are unfortunately also unlikely to be found through training a discriminative ML model. This is because the potentially discriminative features will be deemed not discriminative with the provided labeled weak-supervision data. Even worse, this leads to the problem that data samples appear to have a low uncertainty, as they are considered clear non-matches based on the labeling functions and features used, and are thus not likely to be selected

with priority by the active learning sampling strategy. In practical applications of active learning, this observed behavior results in a quick improvement during the first iterations, i.e., the active learning query strategy is selecting valuable data points, but then to a saturation in which the selection strategy is often not better than random and the trained ML model is only improving marginally.

Diversity sampling-based approaches are in general addressing this issue by trying to sample data points that are more representative of the underlying population, e.g., through:

- *Discovering model-based outliers* by picking data points with low activation in logits and hidden layers.
- *Clustering techniques* to pre-segment the data and per-cluster sampling technique.
- *Representative sampling* by finding items most representative of the target domain.

These approaches work only sub-optimal in matching problems due to the strong imbalance that requires strategies tailored to sampling additional, potential matches. E.g., data points with a low activation will in most cases be non-matches and sampling diversely from different clusters is still mostly resulting in non-matches picks.

It is an object of the present invention to improve and further develop a computer-implemented active learning method in the field of an imbalanced matching situation and a corresponding system for carrying out this method for providing a high degree of performance of the method by simple means.

In accordance with the invention, the aforementioned object is accomplished by a computer-implemented active learning method in the field of an imbalanced matching situation, for finding at least one relationship between data points in different data sets, comprising the following steps:

- executing an uncertainty-based sampling strategy on a defined set of data points;

- 6 -

- performing a detection of a defined saturation level in an improvement of a model performance, wherein the detection is based on outputs of a generative model and a discriminative model;
- further executing the uncertainty-based sampling strategy, if such a saturation level is not detected, and switching to executing a long-tail sampling strategy, if such a saturation level is detected; and
- outputting a final model for prediction, if a defined budget of labels for the data points is depleted.

Further, the aforementioned object is accomplished by a system for carrying out the computer-implemented active learning method in the field of an imbalanced matching situation, for finding at least one relationship between data points in different data sets, comprising:

- executing means for executing an uncertainty-based sampling strategy on a defined set of data points;
- performing means for performing a detection of a defined saturation level in an improvement of a model performance, wherein the detection is based on outputs of a generative model and a discriminative model;
- executing means for further executing the uncertainty-based sampling strategy, if such a saturation level is not detected, and switching means for switching to executing means for executing a long-tail sampling strategy, if such a saturation level is detected; and
- outputting means for outputting a final model for prediction, if a defined budget of labels for the data points is depleted.

According to the invention it has been recognized that it is possible to provide a high degree of performance of the method by using different sampling strategies or methods in a smart way. It has been further recognized that simply performing a detection of a defined saturation level in an improvement of a model performance, wherein the detection is based on outputs of a generative model and a discriminative model, provides a suitable moment in time during performing the whole method for switching from one sampling strategy or method to another sampling strategy or method. Further according to the invention, firstly an uncertainty-based sampling

strategy is executed on a defined set of data points. Then, if such a saturation level is detected, switching from the uncertainty-based sampling strategy to a long-tail sampling strategy is executed. As a result, more potential matches between data points can be found efficiently. At the end of the method a final model for prediction is output, if a defined budget of labels for the data points is depleted.

Thus, the method and system according to the invention provide a high degree of performance of the method by simple means.

According to an embodiment of the invention the uncertainty-based sampling strategy can comprise providing a set X of candidate match pairs of data points from a first source A and a second source B. Such candidate match pairs can for example comprise ontology class pairs or entity pairs.

Within a further embodiment the uncertainty-based sampling strategy can comprise creating of a set of weak supervision signals or functions for a matching problem and outputting a set of weak supervision sources L. Such weak supervision signals can be in the form of weak supervision functions. The matching problem can comprise for example ontology, schema, entity etc.

In a further embodiment the weak supervision signals or functions can be combined in a generative process and at least one label can be created for non-abstained candidate match pairs. The input in this step can be the data pairs X and the set of weak supervision sources L. The output can be predictions and probabilities for non-abstained input data.

According to a further embodiment the at least one label and a set of features can be used to train a discriminative machine learning model to then create predictions for all candidate matches or candidate match pairs or create a temporary classifier. The input in this step can be the input data pairs X and the predictions mentioned in the last paragraph. The output can be predictions and probabilities for all the input data.

Within a further embodiment sampled data points can be provided to an oracle, wherein the oracle can be a computer system or database.

5 In a further embodiment the oracle can provide at least one label for the sampled data points. The oracle can comprise a defined budget of labels.

10 According to a further embodiment the at least one label can be used to re-train a discriminative model or to tune a generative modeling process or to suggest/create new weak supervision sources or signals or functions. The generative modeling process can be tuned through influencing the modelled accuracy of each affected weak supervision source.

15 Within a further embodiment performing a detection of a defined saturation level in an improvement of a model performance can be based on validating the trained model or uncertainty-based sampling strategy with some test dataset or on a defined value of distance between the outputs of the generative model and the discriminative model. In this step the input data X and the predictions and/or probabilities for all the input data can be inputted. The output can be oracle labels/predictions for selected data points.

20 According to a further embodiment the long-tail sampling strategy can comprise computing a feature vector for each data point that needs to be matched. This will contribute to high performance of the method.

25 In a further embodiment the long-tail sampling strategy can comprise a MinHash Locality-Sensitive Hashing, LSH, based similarity computation.

In a further embodiment a MinHash signature can be created for each data point.

30 According to a further embodiment the signature encodes one or more feature vectors. This is a very efficient encoding of feature vectors.

Within a further embodiment the long-tail sampling strategy can comprise computing of informative metric for each data point based on a linear combination of

uncertainty, similarity and weak label availability. This allows for a flexible and adjustable linear combination of information contained in weak labels, model uncertainty and Jaccard similarity.

5 Advantages and aspects of embodiments of the present invention are summarized as follows:

Embodiments can comprise an efficient two-step weakly-supervised active learning method based on the combination of:

- 10
1. The detection of an active learning saturation point for traditional, uncertainty-based sampling.
  2. A long-tail focused sampling strategy tailored for matching problems, to provide more valuable matches in later stages of active learning.

15 Further embodiments can comprise the following steps:

Overall input data: A set of candidate match pairs of data points from Source A and Source B, e.g., ontology class pairs or entity pairs. The set is:  $X = \{(A_1, B_1), (A_1, B_2), (A_2, B_3) \dots\}$ .

20 The method comprising the steps of

- 1) Creation of a set of weak supervision signals, e.g., in form of weak supervision functions, for a matching problem, e.g., ontology, schema, entity...

Output: A set of weak supervision sources  $L = \{L_1, L_2, L_3 \dots L_N\}$ .

- 25
- 2) Combination of these weak signals in a generative process, creating labels for the non-abstained candidate matches.

Input: The input data  $X$  and the set of weak supervision sources  $L$ .

Output: Predictions and probabilities for the non-abstained input data:  $y_{\text{non\_abstained}}$ .

- 30
- 3) Use the labels and a set of features to train a discriminative machine learning model to then created predictions for all the candidate matches.

Input: The input data  $X$  and the predictions  $y_{\text{non\_abstained}}$ .

Output: Predictions and probabilities for all the input data:  $y_{\text{all}}$ .

4) Perform saturation point detection based on outputs of generative and discriminative model and then based on outcome:

a. Perform traditional active learning sampling strategy, e.g., entropy based.

b. Perform long-tail focused sampling strategy.

Input: The input data  $X$  and the predictions  $y_{all}$ .

Output: Oracle labels/predictions for selected data points:  $y_{oracle}$ .

5) Finish when oracle budget is depleted.

6) Output final predictions for matches.

Embodiments of the invention provide a more efficient sampling strategy for imbalanced matching problems. With state-of-the-art methods, one needs to sample and label many more data points to achieve the same performance. This results in more effort/cost. Embodiments of the invention are able to sample more informative data points and thus are able to achieve better matching performance.

An embodiment of the present invention comprises a two-step active learning method for imbalanced matching problems, combining weak-supervision with two subsequent active learning phases, preferably for detecting difficult currently occurring long-tail matches.

A further embodiment of the present invention can be a weakly-supervised active learning method.

Based on embodiments of the present invention more potential matches can be found efficiently, with less time and effort compared with prior art methods. The embodiments provide a key to improve matching performance.

Embodiments of the invention can present a two-step active learning method for imbalanced matching problems, e.g., ontology, schema, entity matching, that can combine weak-supervision with two subsequent active learning phases to detect difficult but commonly occurring long-tail matches. It can use traditional uncertainty-based sampling with performance saturation detection and can then switch to a 2<sup>nd</sup> sampling method that combines traditional uncertainty metrics, an efficient MinHash

LSH based similarity computation, and the available weak supervision signals. Overall, this can achieve among others a better cost-performance trade-off compared to existing methods.

5 There are several ways how to design and further develop the teaching of the present invention in an advantageous way. To this end it is to be referred to the following explanation of examples of embodiments of the invention, illustrated by the drawing. In the drawing

10 Fig. 1 shows in a diagram an overview of an embodiment of the present invention,

Fig. 2 shows an algorithm sketch of an embodiment of a long-tail sampling strategy,

15 Fig. 3 shows a further embodiment wherein matched data is used to create a digital twin and execute optimized building control,

Fig. 4 shows a further embodiment wherein different city data is matched to enable cross-silo optimization applications and

20

Fig. 5 shows a further embodiment comprising creation and deployment of matchings between heterogeneous railroad device infrastructure to enable Smart Railway Management.

25

Embodiments of the invention present a two-phase active learning strategy that addresses the “long tail” problem of finding potential informative matches efficiently. In the first phase, there is executed a traditional sampling strategy until it is detected a saturation point in model performance improvement. Then, it is switched to a second sampling strategy that focuses on these long tail samples using a computationally efficient method, tailored to such matching problems that is based on MinHash LSH based similarity computation, traditional uncertainty metrics, and the available weak supervision signals. Compared to existing works this two-step method achieves higher performance with the same amount of labeling effort. I.e.,

30

less time and effort is needed for annotating data points to achieve equal or even better matching quality.

5 The overall flow of an embodiment of the invention is depicted in Fig. 1. A set of weak-supervision functions creates a set of labels in a generative process for the data points. These labels are then used to train a discriminative model, creating a temporary classifier. The temporary classifier then labels the data. As long as there is labeling budget available, active learning iterations as follows are then performed: First, a saturation detection is performed to decide if default, uncertainty-based  
10 sampling strategies are still sufficient. Then it is executed either such a default sampling strategy, or if saturation has been reached, it is performed a long tail sampling strategy. In either case the sampled data points are provided to the oracle such as a computer system or database, which then will provide labels for the sampled data points. These new, annotated labels can 1.) directly be used to re-  
15 train the discriminative model 2.) to tune the generative modeling process – through influencing the modelled accuracy of each affected weak supervision source – and 3.) an extension to suggest/create new weak-supervision sources.

20 When the defined labeling budget has been depleted, the last trained discriminative classifier is outputted. This classifier is then used as final model for prediction.

In the following two core building blocks, *Saturation Detection* and *Long Tail Sampling Strategy*, of embodiments are described in more detail:

25 **Saturation Detection:**

The simplest way to detect saturation would be to validate the trained model with some test dataset. However, obtaining a representative test data set is costly and active learning is often applied in scenarios where users exactly want to avoid  
30 collecting large, labeled datasets.

There have been developed active learning stopping mechanisms on overall, average uncertainty, based on some uncertainty measure such as entropy or margin uncertainty [1], which is however model dependent and has shown instability, especially in multiclass settings [2].

In embodiments of the present invention, there is proposed a different method. There is exploited the fact that there are two prediction outputs: 1.) the predictions of the generative model and 2.) the predictions of the discriminative model. There is the insight to implement a stopping mechanism based on the overall distances between these two output datasets, i.e., using of sum of squared residuals is possible. Then, the initial sampling strategy is stopped, when the change between successive active learning iterations becomes small enough – defined by some parameter –, and it is switched to the long-tail sampling strategy.

#### Long-Tail Sampling Strategy:

It is proposed the following approach for sampling informative data points for long-tail examples, see also Algorithm 1 for a sketch of an implementation of its core functions:

1. Compute a feature vector for each data point, e.g., a class in an ontology, a data record in a database. Such a feature vector may employ word and sentence embedding models for text attributes or apply data augmentation techniques to augment the raw data.

Output: A feature vector for each data point that needs to be matched.

2. Create MinHash signatures [3] for each data point. These signatures encode the feature vector efficiently. With these MinHash signatures, Jaccard Similarity between combinations in linear time can be quickly estimated:

$$\text{Jaccard Similarity: } J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

3. While oracle budget is available, it is started from a defined similarity\_threshold parameter, e.g., it is started from 1, the highest possible Jaccard similarity):

- a. Create a LSH index with the defined similarity threshold that impacts the chosen number and size of bands used.

- b. Insert all MinHash signatures from dataset A using the data point ID as key.

- c. Query with all signatures from dataset B. This will return all matches above the defined similarity threshold and only requires constant time for each query (thus linear overall).
- d. Filter results by already oracle annotated data points in previous iterations.
- e. Compute informative metric for each data point based on a linear combination of uncertainty, similarity and weak label availability:

$$I(x) = \alpha * |weak_{label}| + \beta * uncertainty + \gamma * jaccard\_similarity$$

This allows for a flexible and adjustable – based on parameters  $\alpha, \beta, \gamma$  – linear combination of information contained in weak labels, model uncertainty and Jaccard similarity.

- f. Sort data points based on computed informative metric.
- g. While oracle budget available:
  - i. Pick batch of highest informative samples to ask the Oracle.
  - ii. If all samples have been provided to the oracle, or the LSH query has not returned any samples, reduce the similarity\_threshold by a defined stepsize parameter. This will loosen the Jaccard similarity restriction and thus return more candidates for active sampling.

### Further Embodiment 1: Match Building Related Data for Digital Twin and Smart Building Applications

Buildings account for roughly 40% of all energy consumed in developed countries. Non-residential buildings, such as office buildings account for a large amount of that energy. These buildings consist usually of several sub systems such as for lighting, heating ventilation and air conditioning, HVAC, access control, room booking etc. For each of this sub-systems, different data schemas/ontologies exist that usually originate in the respective sub-systems and might thus using different naming conventions.

In the architecture, engineering and construction/facility management industry, there exists a common standard for Building Information Models, BIM,

representation, the namely Industry Foundation Classes, IFC, standard. As the BIM represented in IFC contains various in-depth info of the building, including its structure and materials, it can serve as a good basis towards a common data model of a building digital twin. In this digital twin, the detailed IFC information can be used to perform building simulations and as input to optimally actuate the building components. It can also be used to calculate building carbon emissions and develop strategies to reduce these emissions.

The problem is that the underlying data schemas/ontologies are not aligned, and thus a digital twin would need to be constructed for each individual building expensively with lots of man hours. It could be shown that weakly-supervised ontology and schema matching shows great potential for such problems but suffers when labeling functions are not covering the specific matching problems of the building.

On the basis of embodiments of the present invention, active learning can be added effectively, discovering matches beyond the ones covered by the knowledge of the labeling functions. Embodiments of the present invention will accelerate the integration and with little effort enable a digital twin construction and a more optimal operation of the physical building twin.

#### **Further Embodiment 2: Use to match smart city data for Super Smart City Applications**

In cities, data is generated independently by different departments and even private businesses. To create a common data space between these data producers, their non-integrated data needs to be transferred to a common data model. A key challenge here is to provide automated matching on schema and entity level with little available machine learning training data. Enabling such automated matching is greatly accelerated with embodiments of this invention. The then integrated data space can be used for many purposes that affect the city and its operation. E.g., Data from businesses, current events, public transport, and deployed transport sensing infrastructure can be used jointly to optimize traffic flow in the city, e.g., by adjusting traffic lights or deploying more buses.

### **Further Embodiment 3: Use to match data schemas between devices for smart railway management**

5 Smart railways are usually managed with a large amount of heterogenous devices from various manufactures and suppliers. For better system efficiency and automation, those devices very often need to interact with each other either directly or indirectly via some intermediate data processing. To enable such interoperability between devices, domain experts must spend lots of effort and time on mapping  
10 their data schemas. This is a tedious and time-consuming task. Also, it is not a one-time task. As old devices become defect and get replaced with new devices over time and the same device might come from different vendors, domain experts must be always engaged into this data integration loop to ensure the correctness of the data schema mapping. Embodiments of the present invention could help domain  
15 experts to map the data schemas between devices more efficiently and faster.

#### **Extension 1:**

As an extension/alternative use, embodiments of the long-tail sampling strategy can also be used as the only sampling strategy, fully replacing traditional sampling  
20 strategies. This can make sense when either the weak supervision sources are representative and accurate, so that there is no (or not much) room for improvement, or when instead of weak-supervision signals, the ML model is trained with a representative set of labeled of data points.

#### **Extension 2:**

The oracle might also provide new weak-supervision signals or the oracle's annotations might be used to generate new weak-supervision signals or tune existing signals.

30 Many modifications and other embodiments of the invention set forth herein will come to mind to the one skilled in the art to which the invention pertains having the benefit of the teachings presented in the foregoing description and the associated drawings. Therefore, it is to be understood that the invention is not to be limited to the specific embodiments disclosed and that modifications and other embodiments

- 17 -

are intended to be included within the scope of the appended claims. Although specific terms are employed herein, they are used in a generic and descriptive sense only and not for purposes of limitation.

## Claims

1. A computer-implemented active learning method in the field of an imbalanced matching situation, for finding at least one relationship between data points in different data sets, comprising the following steps:

- executing an uncertainty-based sampling strategy on a defined set of data points;
- performing a detection of a defined saturation level in an improvement of a model performance, wherein the detection is based on outputs of a generative model and a discriminative model;
- further executing the uncertainty-based sampling strategy, if such a saturation level is not detected, and switching to executing a long-tail sampling strategy, if such a saturation level is detected; and
- outputting a final model for prediction, if a defined budget of labels for the data points is depleted.

2. The method according to claim 1, wherein the uncertainty-based sampling strategy comprises providing a set of candidate match pairs of data points from a first source (A) and a second source (B).

3. The method according to claim 1 or 2, wherein the uncertainty-based sampling strategy comprises creating of a set of weak supervision signals or functions for a matching problem and outputting a set of weak supervision sources (L).

4. The method according to claim 3, wherein the weak supervision signals or functions are combined in a generative process and at least one label is created for non-abstained candidate match pairs.

5. The method according to claim 4, wherein the at least one label and a set of features is used to train a discriminative machine learning model to then create predictions for all candidate matches or candidate matches or match pairs or create a temporary classifier.

6. The method according to any one of claims 1 to 5, wherein sampled data points are provided to an oracle, wherein the oracle can be a computer system or database.

5

7. The method according to claim 6, wherein the oracle provides at least one label for the sampled data points.

10

8. The method according to claim 7, wherein the at least one label is used to re-train a discriminative model or to tune a generative modeling process or to suggest/create new weak supervision sources or signals or functions.

15

9. The method according to any one of claims 1 to 8, wherein performing a detection of a defined saturation level in an improvement of a model performance is based on validating the trained model or uncertainty-based sampling strategy with some test dataset or on a defined value of distance between the outputs of the generative model and the discriminative model.

20

10. The method according to any one of claims 1 to 9, wherein the long-tail sampling strategy comprises computing a feature vector for each data point that needs to be matched.

25

11. The method according to any one of claims 1 to 10, wherein the long-tail sampling strategy comprises a MinHash Locality-Sensitive Hashing, LSH, based similarity computation.

12. The method according to any one of claims 1 to 11, wherein a MinHash signature is created for each data point.

30

13. The method according to claim 12, wherein the signature encodes one or more feature vectors.

14. The method according to any one of claims 1 to 13, wherein the long-tail sampling strategy comprises computing of informative metric for each data point based on a linear combination of uncertainty, similarity and weak label availability.

5 15. A system for carrying out the computer-implemented active learning method in the field of an imbalanced matching situation, according to any one of claims 1 to 14, for finding at least one relationship between data points in different data sets, comprising:

- 10 - executing means for executing an uncertainty-based sampling strategy on a defined set of data points;
- performing means for performing a detection of a defined saturation level in an improvement of a model performance, wherein the detection is based on outputs of a generative model and a discriminative model;
- 15 - executing means for further executing the uncertainty-based sampling strategy, if such a saturation level is not detected, and switching means for switching to executing means for executing a long-tail sampling strategy, if such a saturation level is detected; and
- outputting means for outputting a final model for prediction, if a defined budget of labels for the data points is depleted.



**Algorithm 1** Long Tail Active Learning Strategy based on MinHash LSH. Sketched building blocks.

---

```

1: budget = 1000                                ▷ Oracle Budget (parameter)
2: threshold = 1.0                                ▷ Jaccard similarity threshold (parameter)
3: stepsize = 0.05                                ▷ Stepsize in which to reduce threshold (parameter)
4: batchSize = 20                                ▷ Number of samples to provide to oracle in one iteration (parameter)
5:
6: function PREPAREMINHASH(datasetA, datasetB)
7:   MinHashesA = dict()
8:   MinHashesB = dict()
9:   for datapoint in datasetA do
10:     feature_vector = computeFeatureVector(datapoint)
11:     MinHashesA[datapoint.id] = computeMinHash(feature_vector)
12:   end for
13:   for datapoint in datasetB do
14:     feature_vector = computeFeatureVector(datapoint)
15:     MinHashesB[datapoint.id] = computeMinHash(feature_vector)
16:   end for
17:   return MinHashesA, MinHashesB
18: end function
19:
20: function GETNEWSAMPLES(MinHashesA, MinHashesB)
21:   combinations = []
22:   while budget > 0 do
23:     if not combinations then
24:       lsh = MinHashLSH(threshold)
25:       for idA, minHashA in MinHashesA do
26:         lsh.insert(idA, minHashA)                ▷ Insert in LSH table using idA as key
27:       end for
28:       for idB, minHashB in MinHashesB do
29:         results = lsh.query(minHashB)              ▷ returns a list of matching idA keys
30:         combinations = combinations + [idA + idB for idA in results]
31:         combinations = filterAnnotatedCombinations(combinations)    ▷ remove already
annotated combinations
32:         combinations = rankCombinations(combinations) ▷ utilizes generative & discriminative
predictions
33:         threshold = threshold - stepsize
34:       end for
35:     end if
36:     queryOracle(combinations[0:batchSize])          ▷ Query with batch of best combinations
37:     combinations = combinations[batchSize:]
38:   end while
39: end function
40:
41: function RANKCOMBINATIONS(combinations)
42:   for c in combinations do
43:     score = a * abs(weak_label) + b * uncertainty + c * jaccard_similarity
44:   end for
45:   return combinations
46: end function

```

---

Fig. 2

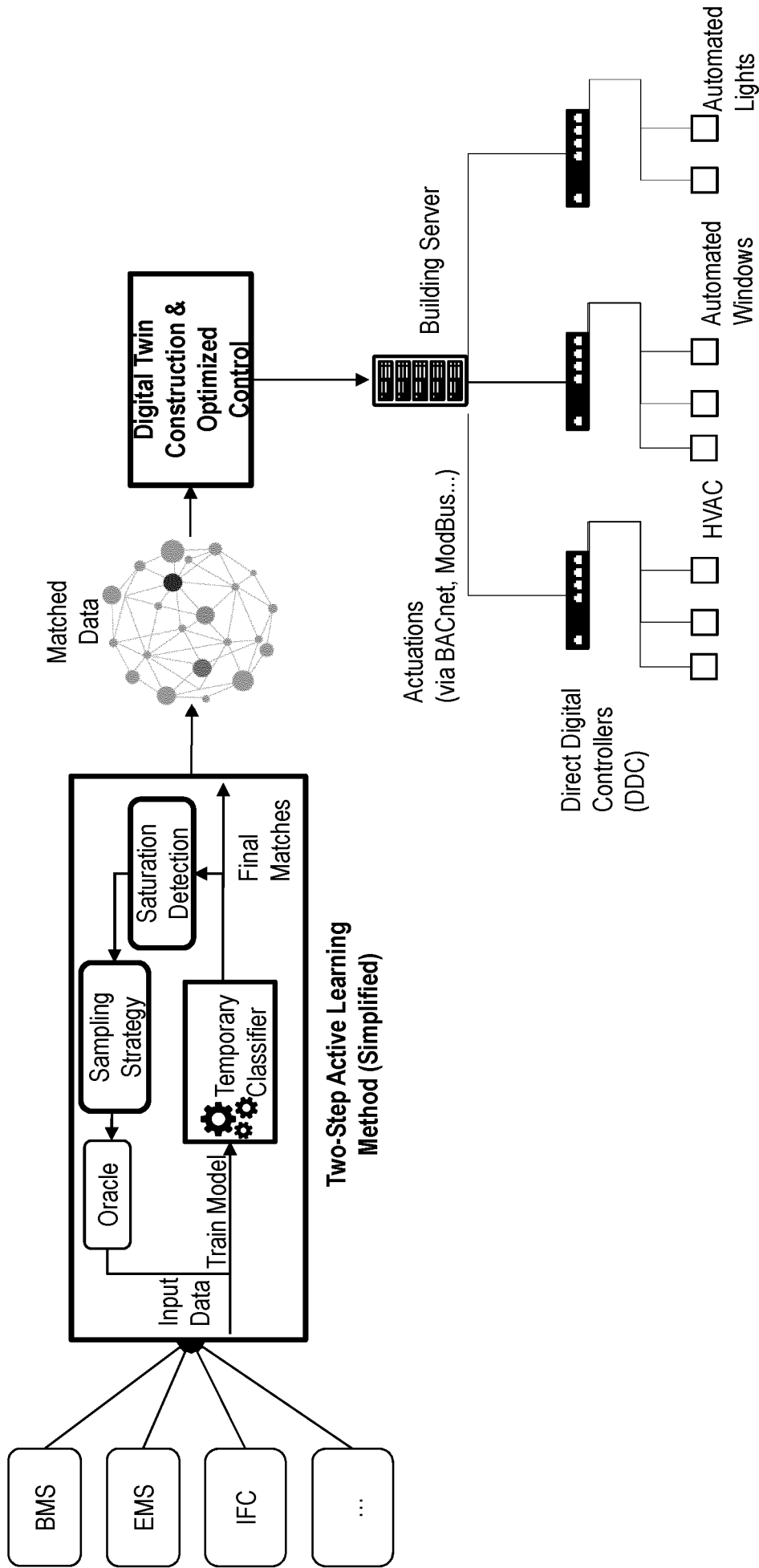


Fig. 3

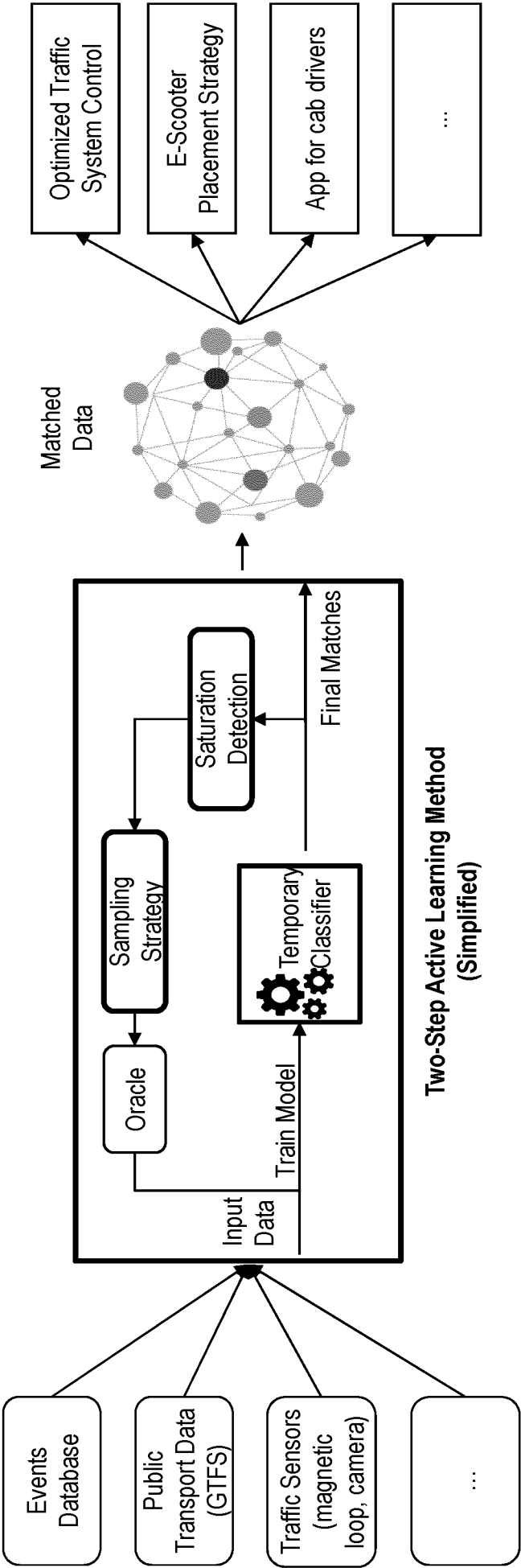


Fig. 4

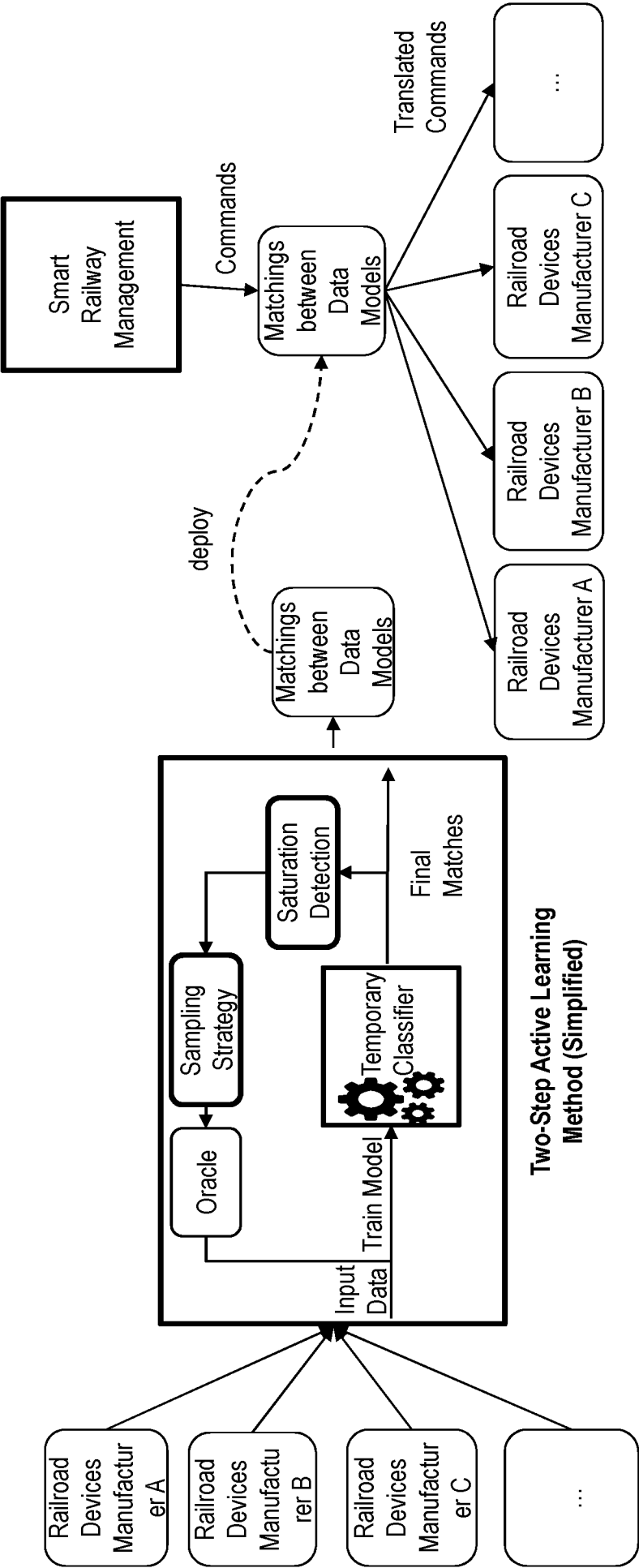


Fig. 5

## INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2022/076440

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                           |                                                    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|
| <b>A. CLASSIFICATION OF SUBJECT MATTER</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                           |                                                    |
| INV.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | G06N3/045                                                                                                                                                                                                                                                                                                                 | G06N3/0475                                         |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                           | G06N3/0895                                         |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                           | G06N3/091                                          |
| ADD.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | G05D1/00                                                                                                                                                                                                                                                                                                                  | G06N5/022                                          |
| According to International Patent Classification (IPC) or to both national classification and IPC                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                           |                                                    |
| <b>B. FIELDS SEARCHED</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                           |                                                    |
| Minimum documentation searched (classification system followed by classification symbols)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                           |                                                    |
| G06N G05D                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                           |                                                    |
| Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                           |                                                    |
| Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                           |                                                    |
| EPO-Internal, WPI Data                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                           |                                                    |
| <b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                           |                                                    |
| Category*                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                        | Relevant to claim No.                              |
| X                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | <b>AJAY J JOSHI ET AL: "Coverage optimized active learning for k - NN classifiers", ROBOTICS AND AUTOMATION (ICRA), 2012 IEEE INTERNATIONAL CONFERENCE ON, IEEE, 14 May 2012 (2012-05-14), pages 5353-5358, XP032194649, DOI: 10.1109/ICRA.2012.6225054 ISBN: 978-1-4673-1403-9 abstract section I, III, VII</b><br>----- | 1-15                                               |
| <input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                           |                                                    |
| * Special categories of cited documents :<br>"A" document defining the general state of the art which is not considered to be of particular relevance<br>"E" earlier application or patent but published on or after the international filing date<br>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)<br>"O" document referring to an oral disclosure, use, exhibition or other means<br>"P" document published prior to the international filing date but later than the priority date claimed<br>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention<br>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone<br>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art<br>"&" document member of the same patent family |                                                                                                                                                                                                                                                                                                                           |                                                    |
| Date of the actual completion of the international search                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                           | Date of mailing of the international search report |
| 13 January 2023                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                           | 27/01/2023                                         |
| Name and mailing address of the ISA/<br>European Patent Office, P.B. 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel. (+31-70) 340-2040,<br>Fax: (+31-70) 340-3016                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                           | Authorized officer<br><br>Papadakis, Georgios      |