Webology (ISSN: 1735-188X)
Volume 18, Number 4, 2021

Human-In-The-Loop Data Classification Framework Using Locality Sensitive Hashing

T. Jebeula[1] , Dr. J. Jebamalar Tamilselvi[2]

[1]Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India.

[2]Associate Professor, Department of Computer Science, SRM Institute of Science and Technology, Ramapuram, Chennai, India.

Abstract—The information generated by the Internet of Things (IoT) is often high dimensional, large volumes with high velocity, and comes from different sources that aggravate the challenges in unifying/pre-processing the data before applying the data analysis to extract the valuable information. Classical ETL framework with rule-based data classification approach may not scale up to handle such high dimensional big data. Machine learning techniques for high dimensional data unification tasks works well. However, it can be accurate only when it does have a sufficient amount of labelled training data. Getting the right labelled data from the big data environment is a challenging task, labor-intensive and cost-prohibitive. This research aims to develop a scalable and efficient data classification framework to classify high dimensional big data by combining probabilistic hashing algorithm (SimHash) and Human-in-the-Loop Machine learning strategy to build a classifier with less labelled training data. Probabilistic hashing algorithms solve the problems associated with the curse of dimensionality in the data classification tasks. The human is brought back into the machine learning loop using active learning strategies through which human, and machine learning processes interact to solve the problem of select the highly informative sample for labelling to build the machine learning model with desired accuracy faster with significantly less labelled data. The evaluation of the proposed data classification framework on a real-time dataset shows that the proposed framework is scalable, sustainable and efficient in classifying high dimensional data.

Keywords—ETL; Active Learning; SimHash; Clustering; Curse of Dimensionality; Uncertainty Sampling; Data Labeling.

I. INTRODUCTION

Internet of Things (IoT) generates massive and inherently high dimensional data. In many cases, the sizes of the datasets have exceeded the memory capacity of a single machine which poses many challenges in pre-processing the data in the data analytics environment. It emphasizes the need for a highly sustainable and scalable data classification framework in the ETL data pre-processing pipeline to efficiently process the high

dimensional data. The classical ETL framework is labor-intensive and prone to errors because of the volume, velocity, variety, and dimensionality of input data. Incorporating Machine Learning algorithms as part of the ETL pipeline for data unification/classification tasks can improve ETL data pre-processing pipeline efficiency. However, it poses additional challenges since it requires sufficient labelled data. This research aims to develop a highly sustainable and scalable high dimensional data classification framework by combining Active Learning (Human-in-the-Loop Machine Learning) strategies with a probabilistic hashing algorithm to accommodate any machine learning model in ETL pre-processing pipeline with less labelled data and computation.

A. Human-in-the-Loop Machine Learning

Active Learning is one of the human-in-the-loop machine learning strategies used to construct a high-performance data classifier while keeping the size of the labelled dataset to a minimum by actively selecting the highly informative record for human annotations [1]. Active Learning uses different sampling strategies to label only a few data samples for training and still achieves high accuracy. The commonly used sampling strategy is uncertainty sampling/posterior probability-based sampling strategies in which the machine learning model selects the sample record from the unlabeled records for which the model is least confident about the prediction or near a decision boundary. Sampling bias is an inherent problem of uncertainty sampling strategy especially applying against the high dimensional big data. These approaches run the risk of selecting outliers and other poor choices of queries. Moreover, selected samples using uncertainty sampling may not be very representative of the input data distribution. Also, chosen records for human annotation may be very similar to each other, which results in collecting more samples for human annotations, which is more challenging in terms of cost, time and effort.

B. Combining Human-in-the-Loop Machine Learning with Locality-Sensitive Hashing Maintaining the Integrity of the Specifications

Locality-Sensitive Hashing algorithms (probabilistic hashing algorithms) are an approximate similarity-based dimensionality reduction algorithm. SimHash is a popular Locality-Sensitive Hashing (LSH) [2] algorithm that hashes similar input [3] text into the same group with the defined high probability. Sim Hash maps a piece of text into a fixed-length fingerprint, and at the same time, the fingerprint has good syntax/semantic consistency. Compressing the data using locality-sensitive hashing algorithms (SimHash) before applying an active learning loop to select the record for human annotations produces a better result in selecting highly informative and diversified samples with less computation and complexity. Active Learning loop can be combined with locality-sensitive hashing algorithms with the following steps.

Step 1: Apply the locality-sensitive hashing algorithms (SimHash) to reduce the dimensionality of the underlying unlabelled data.

Step 2: Structurally cluster the data using the hamming distance, between the hashed text, as the distance measure.

Step 3: Apply the Active Learning loop to select the highly informative and diversified sample for human annotation.

The proposed data classification framework combines both probability hashing algorithm and active learning to build a high-performance, high dimensional data classifier with less labelled data. Also, it improves the efficiency of uncertainty sampling in the active learning loop in selecting diversified and representative sample for human annotations.

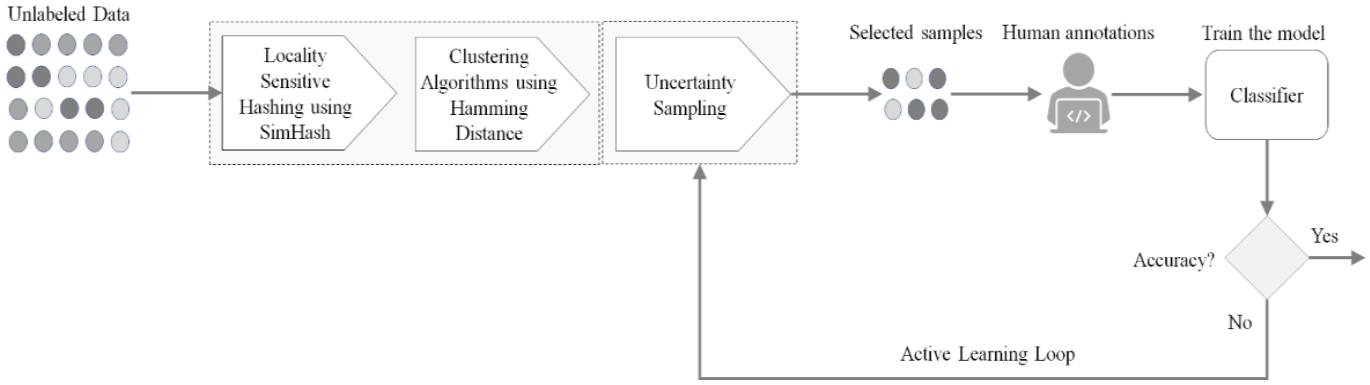

Fig. 1. Proposed Data Classification Framework

The details and summary of this research study are organized as follows. Section II provides a brief survey of Active Learning and dimensionality reduction using probabilistic data structures related to this study. Section III gives the details of crucial components of the proposed data classification framework. Finally, Section IV develops a specific application based on the proposed framework and demonstrates resulting improvements, scalability and classification accuracy.

II. RELATED WORK

In recent years, various probabilistic data structures and Human-in-the-Loop machine learning techniques have been widely studied in various ETL tasks in a big data environment to process the data efficiently with less resource, computation power and high accuracy. Qixia Jiang and Maosong Sun proposed a semi-supervised algorithm based Sim Hash approach for searching similar text in the high dimensional data [4]. Jingbo Zhu et al. developed an active learning solution based on SUD and SBC techniques to address uncertainty sampling problems that frequently fail due to outlier selection. [5]. Tran Ho et al. suggested a new similarity-based method that uses a parallel processing framework to implement the fingerprinting technique [6]. Tianxu He et al. suggested a new active learning framework that takes into account uncertainty, representativeness, and diversity. The suggested approach provides a method for calculating and combining an instance's uncertainty, representativeness, and diversity in a systematic way [7]. Jenq-Haur Wang et al. proposed a clustering algorithm based on Sim Hash to enhance the efficacy in big data analytics using LSH and dimensionality reduction [8]. Zalan Bodo et al. proposed a clustering-based algorithm used in Active Learning, which is based on graph clustering with normalized cuts, and employs a clustering algorithm to extract representative records from the data and approximate spectral clustering to perform the computations more efficiently [9]. The outlier problem of uncertainty sampling was addressed by Zhu et al., who presented two approaches: SUD and density-based re-ranking. [10]. Sudheendra Vijayanarasimhan et al. proposed two hash-based solutions to retrieve near points in sublinear lines [11]. Romaric Pighetti et al.

suggested a new framework that combines supervised learning, multi-objective genetic algorithms, and LSH. They proposed to identify an adapted solution from the initial dataset using Locality Sensitive Hashing (LSH). [12]. Zay Maung Aye proposed a data compression scheme based on LSH for accelerating metric learning. The fundamental idea is to cluster the data using projection-based LSH to choose representative samples of the dataset. [13].

III. PROPOSED FRAMEWORK

This section proposes a data classification framework that combines the Probabilistic Hashing algorithm and Active Learning to build the machine learning-based data classifier model to classify high dimensional data with less labelled data (human-annotated data). The proposed framework has three key components: probabilistic similarity-based data hashing using SimHash algorithm, clustering of data using hamming distance as a distance measure and active learning loop using uncertainty sampling as a sampling strategy to select the label for human annotation. Each key component is explained in the following section.

A. Probabilistic Similarity-Based Hashing using SimHash

Conventional data structures are inefficient for analyzing the massive volume of high dimensional data since the required computational power and time complexity are very high. On the other hand, the probabilistic data structure is quite efficient in reducing the dimensionality of high dimensional data without changing its meaning. Probabilistic hashing algorithms are based on a random hash function that takes the vector representation of the text as the input and returns a value that serves as a fingerprint, which, being discrete, can be used for indexing. These fingerprints generated using probability hashing algorithms are widely used for near-neighbour searching related problems [14]. The performance of LSH largely depends on the underlying particular hashing methods. One of the popular locality-sensitive hashing algorithms [15] is Sim Hash, a sign normal random projection algorithm that is based on the Sim Hash function developed by Moses S.Charikar [16] in 2002 and applied to solve the problem of detecting near-duplicate web pages in Google[17].

Sim Hash is based on the concept of sign random projections. For a N-dimensional document-vector d, it defines a similarity-preserving Sim Hash function for the random vector with components generated from independent and identically distributed normal (i.e.,) as

$$h_v^{Sim}(d) = sign(v.d) = f(x) = \begin{cases} 1, & v.d > 0 \\ 0, & v.d < 0 \end{cases} \qquad (1)$$

Thus, the Sim Hash value generated for the text is the sign of the random projection, and since the hyperplane with a normal vector separates the multidimensional space into two half-spaces, it encodes the information on the side (positive or negative) where the text is located. Thus, if two texts have an angle $\alpha = \pi$, they will appear in different half-spaces, and if text with a perfect alignment that has $\alpha = 0$, it definitely lies in the same half-space. Since the magnitude of the document-vectors does not play any role in formula (1), the probability that two texts d_A and d_B have the same Sim Hash value is equal to the probability of appearing

on the same side of the hyperplane, which can be formulated using the angle between the text as follows which defines the probability of hash collision for the Sim Hash function.

$$P_r\left(h^{sim}(d_A) = h^{sim}(d_B)\right) = 1 - \frac{\alpha}{\pi} \approx \frac{\cos\alpha + 1}{2} \qquad (2)$$

The collision probability between two texts is closely related to the function cos(α); therefore, if the text is close to each other in terms of the cosine similarity, they will almost certainly collide, and vice versa. In this sense, a family of hash functions preserves the cosine similarity between documents and is the locality-sensitive function family for the cosine similarity. The Sim Hash algorithm reduces the dimension of the data which is being analyzed by preserving the similarity. Also, the similarity between texts is then determined with Hamming distance [18]. Thus, Sim Hash has two crucial properties.

(1) The generated fingerprint of a dataset is the hash of dataset's features.

(2) Similar text has similar hash values.

These properties make Sim Hash an ideal technique for grouping approximate similar text into the same bucket [19]. It also improves the data clustering efficiency when applying clustering algorithms on the data after pre-grouping the text using Sim Hash fingerprint, and also, clustering algorithms may not suffer from the curse of dimensionality [20] issues. Furthermore, when using Active Learning on clustered data that has previously been compressed with the Sim Hash technique, the selected sample for human annotations will be highly informative and diverse.

B. Pre-Cluster the data using Hamming Distance

The hamming distance between two vectors and corresponding to the string and is defined as the number of bit positions where and differ. Hamming distance is the criterion to judge the similarity of the SimHash of two different text [21]; the smaller the Hamming distance is, the greater the similarity. Therefore, Pre-cluster the hashed data (hashed using Sim Hash algorithm) using a clustering algorithm using hamming distance as the distance measure improves the efficiency of the active learning loop in selecting informative and diverse sampling.

Algorithm: Dimensionality reduction and pre-clustering of unlabeled data.

U←Unlabeled data.

E_{dset1}, E_{dset2}, E_{dset3}←Empty dataset.

//Arrive simhash fingerprint for each unlabeled record.

for each unlabeled record r in U

begin

f_p←simhash_signature(r)

append the dataset $E_{dset1}$with (r, f_p)

end

//Dimensionality reduction

for each unlabeled record r in $E_{dset1}$with simhash fingerprint f_p

begin

append the dataset $E_{dset2}$with (r, f_p) if f_p not found in E_{dset2}

end

//Clustering (Structural data grouping)

E_{dset3}← cluster_dataset(dataset: E_{dset2}, distance_measure: hamming_distance)

return clustered dataset E_{dset3}

C. Active Learning Loop using Uncertainty Sampling

Uncertainty Sampling is an active learning sampling strategy to identify highly informative unlabeled data that are near the decision boundary of the specific selected Machine Learning model [22]. Following are the four types of commonly used uncertainty sampling techniques used to select the more informative sample record from the unlabeled dataset for human labelling.

1) Least Confidence: Select the record for which the model/learner is least confident about the prediction [23].

2) Margin of Confidence: Select the record with a very less difference between the first and second most likely predictions [24].

3) Ratio of Confidence: Select the record with a very less ratio difference between the first and second most likely prediction.

4) Entropy: Select the record based on the record's average information content [25].

In this study, the Least Confidence sampling strategy is used. For the given predicted probability of all the labels in the dataset, the following equation (3) is the probability of the highest confidence y* for the label.

$$\phi_{LC}(x) = P_\theta(Y^*|X) \qquad (3)$$

Rank the predicted probability order by the confidence level and convert the same into the scores of 0 to 1 range. 1 indicates the most uncertain score. The sampling based on least confidence with 0 to 1 range is defined as (4)

$$\phi_{LC}(x) = (1 - P_\theta(Y^*|X)) * \frac{n}{n-1} \qquad (4)$$

IV. PROPOSED DATA CLASSIFICATION FRAMEWORK

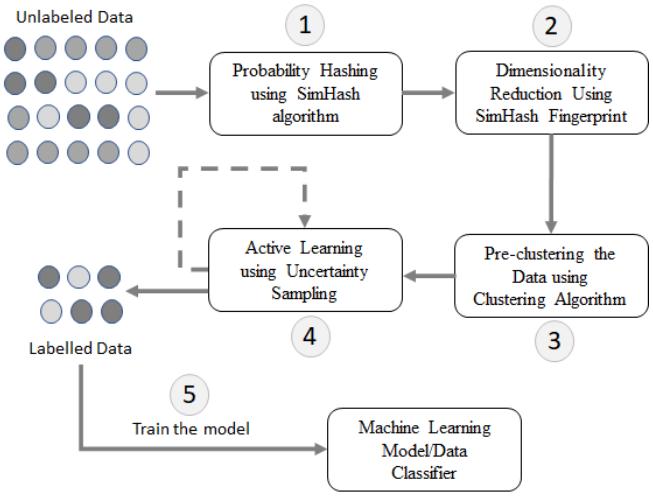

Fig. 2. Proposed Data Classification Framework using SimHash and Active Learning

In the proposed framework, Active Learning is combined with a probabilistic hashing algorithm (Sim Hash) to create a data classification model that can categorize high dimensional data with considerable less labelled training data. The proposed framework is scalable, sustainable and effective at classifying high dimensional source data with less labelled training data. Unlabeled data is getting hashed using the Sim Hash algorithm as a first step in the proposed framework. After removing all duplicate records (retaining unique records alone) based on the Sim Hash fingerprint, data is again clustered using a clustering algorithm with hamming distance as the distance measure. On top of clustered data, active learning with the least confidence sampling strategy is applied to select the informative and diversified sample records from each cluster for human annotations. After the human annotations are completed, the training dataset is updated with annotated labels, and the model is trained with the new training dataset. Based on the model's accuracy, the active learning loop keeps getting applied till it reaches the desired accuracy level. Following is the algorithm outline of the proposed data classification framework.

- simhash_signature (r): receives unlabeled record r as the input and generates the Sim Hash signature f_p by preserving the similarity among the data in the unlabeled dataset.
- cluster_data(dataset, distance_measure): receives the unlabeled data set with the corresponding Sim Hash signature f_p for each record and distance measure as the input and clusters the data based on the provided distance measure.
- uncertainty_sampling (cluster,classifier,sampling_technique): receives the unlabeled data set, classifier and sampling strategy[26] as the input and sample record according to the sampling strategy.
- get_human_annotations(data samples): receives the sampled data set as the input and get the human annotation/labelling.

Algorithm: Proposed algorithm using SimHash and Active Learning

C←Data Classifier
T_{limit}←Classifier accuracy threshold limit
L← Labelled data set
U← Unlabeled data set
E_{dset1}, E_{dset2}, E_{dset3}← Empty data set

train the classifier C with labelled data set L
//Arrive simhash fingerprint for each record.
for each unlabeled record r in U
begin
    f_p←simhash_signature(r)
    append the dataset E_{dset1} with (r, f_p)
end
//Dimensionality reduction
for each record r in E_{dset1} with simhash fingerprint f_p
begin
    append dataset E_{dset2} with (r, f_p) if f_p not found in E_{dset2}
end
//Clustering (Structural data grouping)
E_{dset3}←cluster_dataset(dataset: E_{dset2}, distance_measure: hamming_distance)

step A:
//Active learning loop
for each cluster c in E_{dset3}
begin
    S←uncertainty_sampling (cluster:c, classifier: C, sampling_technique: least_confidence_sampling)
    L←get_human_annotations (data samples:S)
End
//Re-train the model
train the classifier C with updated labelled data set L
if accuracy of the classifier C > T_{limit}
then
    end active learning loop
else
    repeat the step A
end

V. EXPERIMENTAL RESUTLS

The efficiency and scalability of the proposed data classification framework have been validated on "Million News Headlines" dataset posted on Kaggle. The chosen dataset and its properties are summarized in Table I.

TABLE I. **MILLION NEWS HEADLINES DATASET**

Context	No. of columns	Total Unique Values
It contains data collected from news headlines over a period of eighteen years.	2	1195191

This experiment aims to validate the efficiency and scalability of the proposed data classification framework and observe the classifier's accuracy against the number of labels selected for human annotations from the high dimensional input data. Also, compare the efficiency and accuracy of the proposed framework against active learning with a cluster-based diversity sampling strategy. The number of records considered for the Active Learning loop with respect to the proposed framework has been detailed in Table II.

The proposed framework outperforms the cluster-based active learning strategy with respect to selecting a highly informative sample for human annotations from the high dimensional data. Furthermore, it has been observed that the proposed framework is more scalable and performance efficient in classifying high dimensional sparse data.

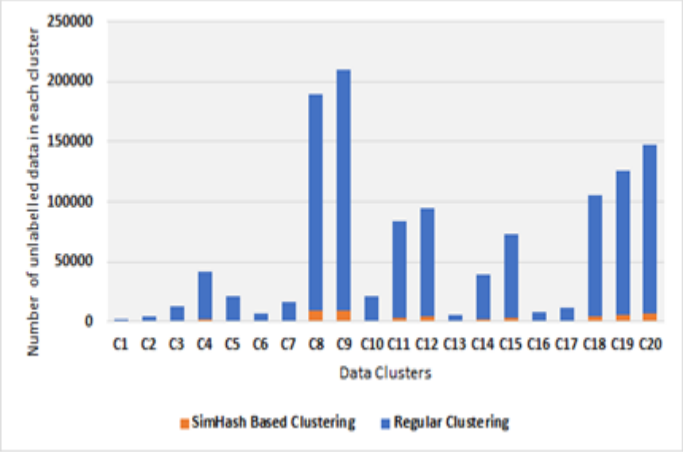

Fig. 3. **Sampling Space after Probability Hashing**

Figure 3 shows the percentage of records considered for active learning loop after probability hashing using Sim Hash. After dimensionality reduction using a probability hashing algorithm, the sampling space (the records to be considered for the active learning loop) is reduced to 5% approximately. Thus, the

subsequent active learning loop to select the label for human annotation is more efficient in selecting a highly informative and diversified sample with less computational power due to the similarity preserving dimensionality reduction on the original dataset. Total record in the initial dataset and record considered for active learning loop after probability hashing is as follows.

- Total unique records: 1195191
- Total records considered for active learning loop after probability hashing: 58200

TABLE II. SAMPLING SPACE AFTER PROBABILITY HASHING

S. No.	Regular Clustering	Sim hash Based Clustering
1	854	100
2	4389	200
3	222339	600
4	142661	2000
5	110231	1000
6	17643	350
7	85928	800
8	39772	9000
9	38043	10000
10	229681	1000
11	1261	4000
12	88586	4500
13	19024	300
14	19307	1900
15	4963	3500
16	6645	400
17	656	550
18	23512	5000
19	5077	6000
20	2226	7000

Table II shows the number of records in each cluster before applying probability hashing and after applying probability hashing. It has been evident that Sim Hash algorithm reduces the data dimension drastically, which can accelerate the efficiency of active learning in selecting the highly informative and diversified sample when pre-clustering the data further using hamming distance. Figure 4 shows that classification accuracy vs number of labels selected for human annotations.

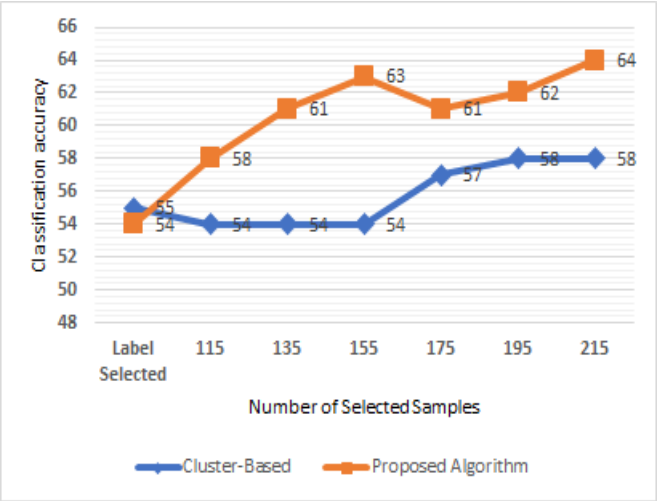

Fig. 4. Selected Samples Vs Accuracy

VI. CONCLUSION

This paper proposed a machine learning-based data classification framework that combines active learning and probabilistic hashing algorithms (Sim Hash). The proposed data classification framework in the ETL data pre-processing pipeline can efficiently classify high dimensional data with fewer training records. Also, it can adopt any machine learning model in classifying high dimensional data since removing the curse of dimensionality issues is the inherent feature of the framework. The scalability and accuracy of the framework have been validated against the real-time data and confirmed that the proposed data classification framework is computationally efficient in classifying high dimensional data with less labelled training data. Furthermore, it has been observed that the proposed framework accelerates the active learning loop in terms of selecting the informative sample for human annotation in the high dimensional data space.

REFERENCES

[1] F. Breve, “Combined active and semi-supervised learning using particle walking temporal dynamics” , Proc. - 1st BRICS Ctries. Congr. Comput. Intell. BRICS-CCI 2013, pp. 15–20. https://doi.org/10.1109/BRICS-CCI-CBIC.2013.14

[2] Y. M. Kwon, J. J. An, M. J. Lim, S. Cho, and W. M. Gal, “Malware classification using simhash encoding and PCA (MCSP)”, Symmetry (Basel)., 2020, vol. 12, no. 5, pp. 1–12. DOI:10.3390/sym12050830

[3] C. Sadowski and G. Levin, “SimHash : Hash-based Similarity Detection, Techreport”, 2007, pp. 1–10.

[4] Q. Jiang and M. Sun, “Semi-Supervised SimHash for efficient document similarity search,” ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., vol. 1, pp. 93–101, 2011.

[5] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, “Active learning with sampling by uncertainty and density for word sense disambiguation and text classification”, Coling 2008 - 22nd Int. Conf. Comput. Linguist. Proc. Conf., vol. 1, no. August, pp. 1137–1144, 2008.

[6] P. T. Ho, H. S. Kim, and S. R. Kim, “ Application of sim-hash algorithm and big data analysis in spam email detection system”, Proc. 2014 Res. Adapt. Converg. Syst. RACS 2014, pp. 242–246, 2014.

[7] T. He et al., “An active learning approach with uncertainty, representativeness, and diversity”, Sci. World J., vol. 2014.

[8] J. Wang and J. Lin. “ K-Means Algorithm for Big Data Analytics”, pp. 1881–1888, 2016.

[9] J. Jiang and H. H. S. Ip. “Active Learning with SVM”, Encycl. Artif. Intell..2019.

[10] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, “Active learning with sampling by uncertainty and density for data annotations”, IEEE Trans. Audio, Speech Lang. Process., vol. 18, no. 6, 2010, pp. 1323–1331.

[11] S. Vijayanarasimhan, P. Jain, and K. Grauman. “Hashing hyperplane queries to near points with applications to large-scale active learning”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 2, 2014, pp. 276–288.

[12] R. Pighetti, D. Pallez, and F. Precioso. “Improving SVM training sample selection using multi-objective evolutionary algorithm and LSH, Proc. - 2015 IEEE Symp. Ser. Comput. Intell. SSCI 2015, pp. 1383–1390.

[13] Z. Maung and M. Aye. “Scaling Learning Algorithms using Locality Sensitive Hashing”, no. June, 2018, p. 20.

[14] O. Jafari, P. Maurya, P. Nagarkar, K. M. Islam, and C. Crushev. “A Survey on Locality Sensitive Hashing Algorithms and their Applications”, ACM Comput. Surv.2021.

[15] A. Wylie, “Locality-Sensitive Hashing”, 2013.

[16] M. S. Charikar. “Similarity estimation techniques from rounding algorithms”, Conf. Proc. Annu. ACM Symp. Theory Comput., 2002, pp. 380–388.

[17] G. S. Manku, A. Jain, and A. Das Sarma. “Detecting near-duplicates for web crawling”, 16th Int. World Wide Web Conf. WWW2007, pp. 141–150.

[18] X. Feng, H. Jin, R. Zheng, and L. Zhu. “Near-duplicate detection using GPU-based simhash scheme”, Proc. 2014 Int. Conf. Smart Comput. pp. 223–228.

[19] O. Jing. “Research on the Fast Retrieval Algorithm of English Sentences Based on Simhash”, Proc. 2020 IEEE Int. Conf. Power, Intell. Comput. Syst. ICPICS 2020, pp. 997–1000, 2020.

[20] M. Steinbach, L. Ertöz, and V. Kumar. “The Challenges of Clustering High Dimensional Data” , New Dir. Stat. Phys.,2004, pp. 273–309.

[21] Y. Yuan, R. Li, Y. Wang, T. Cao, J. Yang, and Y. La. “Application of the maintenance data of transformers based on SimHash and Hamming distance algorithm”, 7th IEEE Int. Conf. High Volt. Eng. Appl. ICHVE 2020 - Proc., pp. 1–4.

[22] D. D. Lewis and W. A. Gale. “A sequential algorithm for training text classifiers”, Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 1994, pp. 3–12.

[23] A. Culotta and A. McCallum. “Reducing labeling effort for structured prediction tasks”, Proc. Natl. Conf. Artif. Intell., vol. 2, 2005, pp. 746–751.

[24] T. Scheffer, C. Decomain, and S. Wrobel. “Active hidden markov models for information extraction”, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2189, 2011, pp. 309–318.

[25] I. Dagan, S. P. Engelson, and R. Gan. “Committee-Based Sampling For Training Probabilistic Classifiers,” 1993.

[26] M. Elgendy. “Human-in-the-Loop Machine Learning”, Version 1 MEAP Edition Manning Early Access Program Copyright 2019 Manning Publications.