Machine Learning Models for Automatic Labeling: A Systematic Literature Review

Teodor Fredriksson[1], Jan Bosch[1][a] and Helena Holmstrm Olsson[2]

[1]*Department of Computer Science and Engineering, Division of Software Engineering,*
Chalmers University of Technology, Gothenburg, Sweden
[2]*Department of Computer Science and Media Technology, Malm University, Malm, Sweden*

Keywords: Semi-supervised Learning, Active Machine Learning, Automatic Labeling.

Abstract: Automatic labeling is a type of classification problem. Classification has been studied with the help of statistical methods for a long time. With the explosion of new better computer processing units (CPUs) and graphical processing units (GPUs) the interest in machine learning has grown exponentially and we can use both statistical learning algorithms as well as deep neural networks (DNNs) to solve the classification tasks. Classification is a supervised machine learning problem and there exists a large amount of methodology for performing such task. However, it is very rare in industrial applications that data is fully labeled which is why we need good methodology to obtain error-free labels. The purpose of this paper is to examine the current literature on how to perform labeling using ML, we will compare these models in terms of popularity and on what datatypes they are used on. We performed a systematic literature review of empirical studies for machine learning for labeling. We identified 43 primary studies relevant to our search. From this we were able to determine the most common machine learning models for labeling. Lack of unlabeled instances is a major problem for industry as supervised learning is the most widely used. Obtaining labels is costly in terms of labor and financial costs. Based on our findings in this review we present alternate ways for labeling data for use in supervised learning tasks.

1 INTRODUCTION

In software-intensive companies in the online and in the embedded systems domain huge sets of data are being processed and labeled manually, either by one or several of their employees (AzatiSoftware, 2019). This is an expensive approach for a company but it does allow for easy maintenance of the quality of the data. The only downside is that the task will be tedious and time consuming and prohibitively expensive due to the human factor.

Data labeling is a way of annotating data depending on the content of the data (see (AzatiSoftware, 2019)). The labels each data entry receives is decided after information about the entry has been processed.

The modern and most reasonable way to try and perform every task nowadays is to use artificial intelligence and for automatic labeling there is no difference.

There are three different ways that machine learning algorithms learn, *Reinforcement learning*, *Supervised learning* and *Unsupervised learning*. Out of the three types of learning, none of these can fully solve the labeling problem.

There are two disciplines in machine learning that are designed for the sole purpose of using data that is either unlabeled or contains a small set of labeled instances. These two labeling methods are called *active learning* and *semi-supervised learning*. For the remainder of this paper we will systematically map the research that has been conducted towards semi-supervised and active learning techniques and how they are applied to labeling. We are particularly interested in how these methods can be applied in different industrial scenarios and therefore we will categorize all the possible research areas. We present this work in hope that will contribute in inspiring people in the industry to use active and semi-supervised learning techniques for labeling tasks.

The contribution of this paper is threefold. First, we provide an overview of the available approaches for (semi-)automatic labeling of data for machine learning based on a systematic literature review. Sec-

[a] https://orcid.org/0000-0003-2854-722X

Fredriksson, T., Bosch, J. and Olsson, H.
Machine Learning Models for Automatic Labeling: A Systematic Literature Review.
DOI: 10.5220/0009972705520561
In *Proceedings of the 15th International Conference on Software Technologies (ICSOFT 2020)*, pages 552-561
ISBN: 978-989-758-443-5

ond, we present the data types that are typically subject to (semi-)automatic labeling and which data types require additional research. Finally, we identify the open research questions that need to be addressed by the research community.

The remainder of this paper is organized as follows. In the next section, we provide the background and an overview of techniques and approaches for automatic labeling. In section 3, we provide a concise description of the problem that we seek to address in this paper followed by an overview of our research method in section 4. We present the results of the systematic literature review in section 5 and discuss these in section 6. Finally, we end the paper with an overview of open research questions in section 7 and a conclusion in section 8.

2 BACKGROUND

As mentioned above, most machine learning paradigms are either supervised or unsupervised. This means that we have access to labels, or we do not have access to labels. Note that most algorithms in the industry are supervised but most data do not have labels and so we need additional efforts to produce good labels. Furthermore, it is not unreasonable to think that a small subset of each big dataset coming from companies have labels and this is where semi-supervised learning is applicable, together with some of the active learning framework.

2.1 Semi-supervised Learning

Semi-supervised learning is a set of machine learning algorithms that can be used if most of the instances are unlabeled, but a small subset of them have labels. In technical terms, we have access to a set of data points that can be divided into two disjoint subsets, one containing the labeled instances and the other containing the unlabeled instances. The objective of semi-supervised classification is to train a classifier on both unlabeled and labeled data so that it is better than a supervised classifier trained only on the labeled data.

Areas of semi-supervised learning can be found in (Zhu and Goldberg, 2009):

1. *(Generative) Mixture Models and EM Algorithm.*
2. *Co-training and Multi-view Learning.*
3. *Graph-based Semi-supervised Learning.*
4. *Semi-supervised Support Vector Machines (S3VM).*

2.2 Active Learning

Historically machine learning algorithms usually try to fit a model according to currently labeled data and we refer to these models as ”passive” learning models. Active learning systems on the other hand creates new models as it iterative learns. Similar to how a scientist plans several experiments to help come to a conclusion about a hypothesis, an active learning method imposes query strategies to help select the most informative examples to be labeled by an oracle.

In some cases, e.g the model does not require a huge number of labels, an active learning system might not be optimal. Instead use it when there is a very big set of unlabeled examples and you need to label a huge amount of data to train the system.

If active learning is appropriate for your problem, then we need to specify in what way we want to query the examples (Settles, 2012). The three most common scenarios are:

1. *Query synthesis*
2. *Stream-based selective sampling*
3. *Pool-based sampling*

As presented in (Settles, 2012), areas of active learning queries include:

1. *Uncertainty Sampling.*
2. *Query by Committee/Disagreement (QBC/QBD).*
3. *Expected Error/Variance Reduction.*

3 PROBLEM STATEMENT

Data labeling is an essential step when pre-processing data to use with machine learning when preforming supervised learning since it is dependent of the presence of labels.

According to reports, up to 80% (see (CloudFactory.com, 2019)) of the time that companies spent on their machine learning projects are allocated to do task such as cleaning, pre-processing and labeling data which is valuable time spent doing other tasks. For example, an ML system that is trained to recognize different animal species in a picture needs training data that contains images that already have labels. Another good example is autonomous vehicles such as self-driving cars. These cars are not safe enough to be deployed in traffic. In order for them to be safer they need to able to distinguish between different objects in its path. Therefore, we need to train the AI of the car using images where the key features are labeled.

3.1 How does Data Labeling Work?

ML systems uses large datasets for training in order to develop a strong AI that can learn patterns. This training data must be labeled or annotated based on the most essential features so that the model can organize the data in the best possible way.

It is essential to use labels that are informative and independent to create an algorithm of top quality. A well labeled dataset provides a ML model with empirical evidence to evaluate the accuracy of the model. The model is then refined.

A ”quality algorithm” is an algorithm that has both high ”accuracy” and high ”quality”, where ”accuracy” refers to how good the predicted labels are and ”quality” refers to how consistent the dataset is.

Errors in the data labeling will worsen the quality of the training data and so the performance of any models used for prediction. To avoid these errors several organizations chooses to implement HITL (Human-in-the-loop) so that to keep humans involved in the training and testing of the models through the deployment phase. HITL is studied within *Interactive Machine Learning* (iML) (Holzinger, 2016).

3.2 Methods for Labeling

Companies have several different ways they can acquire labels for their data, popular choices are:

- **Crowd-sourcing:** Allows companies to preform labeling more quickly by having access to a lot of people and divide the labeling task among these people rather than just using one employee for the job.(CloudFactory.com, 2019)
- **Contractors:** Companies employ outside freelancers temporarily for labeling (CloudFactory.com, 2019).
- **Managed Teams:** Companies gives the labeling task to a group that they train specifically for labeling, this team is usually managed by a third-party organization (CloudFactory.com, 2019).
- **In-house Staff:** The company enlists the labeling the current employees. (CloudFactory.com, 2019)

There is no definite way of labeling data optimally and companies have to decide by themselves on how their labeling should be done. When selecting a data labeling method, the main factors are the following:

1. Financial costs.
2. The size of the dataset.
3. The knowledge of the staff.
4. What is the objective of the ML model that needs labels.

The team performing the labeling must have an excellent knowledge of the industry and its servers, they need to be flexible since labeling and machine learning is an ever-changing process that is quickly evolving as more data is coming in.

4 RESEARCH METHOD

This section presents the research method used in the study, namely systematic literature review. Systematic literature review seeks to identify, analyse, and interpret all relevant research (i.e., primary studies) on the topic of interest (Keele et al., 2007). In this study the topic of interest is data labeling in machine learning and the goal our SLR is to identify and analyze literature in this research area. We followed the procedure of conducting systematic literature reviews according to (Keele et al., 2007). The procedure can be summarized as follows:

1. Definition of research questions.
2. Identification of search terms and conducting search.
3. Screening of papers on the basis of inclusion and exclusion.
4. Data extraction and mapping.

For the rest of this chapter we will outline this procedure.

4.1 Definition of Research Questions

The purpose of this study was to establish what current research has been accomplished in the field of automatic labeling of data from different fields using different machine learning method. Thus the main objectives of this literature review is:

- Examine previous research on the subject of automatic labeling.
- Explore the possibility of contributing with new research within the area.

We define a number of research questions and their motivations below.

RQ1. In what research fields can we apply active and semi-supervised learning?: *This RQ seeks to identify different research fields that exploit active and semi-supervised learning*

RQ2. What kind of machine learning algorithms are used?: *This RQ seeks to identify what type of different active learning and semi-supervised learning paradigms can be used.*

RQ3. What is the popularity of data types among the different methods?: *This RQ seeks to identify for each method, we how many papers studied a specific datatype.*

4.2 Identification of Studies

Keyword-based database search was used to source relevant studies.

In this study, the main search string that was constructed iteratively consisted of the two keywords: Active machine learning OR semi-supervised learning. First we performed pilots with other keywords, such as ”automatic labeling” but it gave a too wide range of methods that were hard to categorize. This was then changed to active learning and semi-supervised learning methods as they were easier to categorize

We further went on to improve the search string in the following way. First, for active learning we searched for ”active machine learning” + ”category of active learning”. If we dismissed the ”machine” in the string, we would get results related to ”education”. Similarly for semi-supervised learning” we searched for ”semi-supervised” learning ” + ”category of semi-supervised learning”. The categories can be located in table 2. Some methods that are not included in this study that are being researched are:

- Constrained clustering.(Basu et al., 2008), (Brefeld et al., 2006)
- Semi-supervised regression.(Cortes and Mohri, 2007), (Sindhwani et al., 2005),(Zhou and Li, 2005).
- Model and feature selection using unlabeled data. (Kääriäinen, 2005), (Madani et al., 2005), (Schuurmans and Southey, 2002), (Li and Guan, 2008).
- Label sampling such as multi-instance learning, multi-task learning and deep learning.(Rosset et al., 2005). (Zhou and Xu, 2007), (Liu et al., 2008), (Ranzato and Szummer, 2008), (Weston et al., 2012).

We did not directly include any of these in the search string as we could not find any relevant papers containing any industrial application.

The search string was applied to Google Scholar. Since the search terms are so general we expect a large number of relevant articles from the search so we deem it sufficient only to use Google Scholar (https://scholar.google.com). The second reason to only use Google scholar is because Google scholar is perceived as an unbiased source according to (Wohlin, 2014). Furthermore, we do not limit ourselves to any time period since the rise of machine learning computations was from around the year 2000 and papers between 1980 and 1999 should mostly contain theoretical research that we deem unnecessary for our study purpose.

The search strings were applied in December 2019 to the selected electronic database to retrieve articles that include the keywords in their title, abstracts and instructions. To avoid ending up with an infinite amount of papers the retrieval stopped after the abstracts and introductions became less relevant. At the end, approximately 300 articles were retrieved for further screening and processing of inclusion and exclusion criteria.

4.3 Study Selection: Inclusion and Exclusion Criteria

All retrieved studies were examined for inclusion and exclusion based on pre-established criteria. The exclusion and inclusion criteria considered in our study are presented below:

Inclusion Criteria.

- Papers that includes AL/SSL techniques for labeling unlabeled and or partially unlabeled data form the industry.
- Papers that compare several AL/SSL techniques with each other.
- Papers that include a hybrid between AL/SSL learning.
- Papers that compare AL/SSL techniques with other non-AL/SSL methods.
- Papers that has a title that describes the application.

Exclusion Criteria.

- Papers concerning theoretical proofs of AL/SSL methods.
- Papers concerning simulation studies.

4.4 Data Extraction and Analysis

Data extraction involved the collection of information related to the RQs of the study. For each paper we identified the research field, what kind of datatype it was and what method the paper focused on.

4.5 Threats to Validity

Although we did not include deep learning in our search string some papers might include deep learning because active or semi-supervised learning was applied to a deep neural network. Some of the papers will contain theoretical properties as well as empirical evaluation of the models. There is no way of telling whether the data sets used in the papers have been tampered with to fit the models better.

5 RESULTS

In section we will interpret the results that we gathered based the research questions in the previous section.

5.1 RQ1: In What Research Fields Can We Apply Active and Semi-supervised Learning?

Table 1 shows how we categorize the different types of data. Going from left to right, the first column contains the names of each category, the second columns shows which datatypes belong to that category, the third columns says what research areas are covered in each category and the last column references each paper that was used for each category.

5.2 RQ2: What Kind of Machine Learning Algorithms Are used?

In this subsection we present the main active and semi-supervised machine learning approaches based on textbooks (Settles, 2012), (Zhu and Goldberg, 2009).

Table 2 shows a summary of the popular machine learning methods for labeling (Settles, 2012), (Zhu and Goldberg, 2009). In the left column we see the active learning methods (Settles, 2012) and in the right column we see the semi-supervised learning methods (Zhu and Goldberg, 2009).

Table 3 shows how we have categorized the semi-supervised learning methods. In the left column we see the name of each category and in the right column we see what method(s) below to each category. We did not include cluster based active learning and cluster-then-label semi-supervised learning in the search string as we did not find any papers relevant to industry.

Figure 1 shows an overview that illustrates how many papers focused on each of the active learning and semi-supervised learning methods. On the horizontal axis we have the methods and on the vertical axis we have the number of papers that focused on that particular paper. The most popular category is co-training and multi-view learning with a total of eleven papers (Yan and Naphade, 2005), (Morsillo et al., 2009), (Zhang and Zheng, 2017),(Di and Crawford, 2011),(Guan et al., 2007), (Rigutini et al., 2005), (Guo and Xiao, 2012), (Cui et al., 2011), (Yu et al., 2010b), (Wu et al., 2019), (Jing et al., 2017), second place is shared with graph-based semi-supervised learning (Tang et al., 2009), (Tang et al., 2011), (Tang et al., 2008), (Abbasi et al., 2015),(Zhao et al., 2015), (Liu and Kirchhoff, 2013), (Zeng et al., 2013), (Stikic et al., 2009) (Chen et al., 2008) and uncertainty sampling (Liu et al., 2016), (Rajan et al., 2008), (Minakawa et al., 2013), (Yu et al., 2010a), (Zhu et al., 2009), (Zhu et al., 2008), (Zhang and Chen, 2002), (Varadarajan et al., 2009), (Kim et al., 2006). (Colares et al., 2013), (Shi et al., 2010), (Huang and Hasegawa-Johnson, 2009), (Nigam et al., 2006).

Co-training and multi-view methods corresponds to 25.00% of all the methods. Graph-based methods and uncertainty sampling both corresponds to 20.45% each. Last but not least mixture models lands at fourth place with 9.09%.

5.3 RQ3: What is the Popularity of Datatypes among the Different Methods

Here we only present graphs for the most popular methods graph-based, co-training, multi-view learning, mixture models and uncertainty sampling. The rest are omitted due to insufficient amount of data.

The first plot from the left of figure 2 illustrates that for multidimensional inputs we found four relevant papers (Tang et al., 2009), (Tang et al., 2011), (Tang et al., 2008), (Abbasi et al., 2015), for sequential inputs we found four relevant papers (Zhao et al., 2015), (Liu and Kirchhoff, 2013), (Stikic et al., 2009), (Stikic et al., 2009) and no relevant one paper for univariate inputs (Chen et al., 2008).

The second plot from the left of figure 2 illustrates that for multidimensional inputs we found four relevant papers (Yan and Naphade, 2005), (Morsillo et al., 2009), (Zhang and Zheng, 2017),(Di and Crawford, 2011). For sequential inputs we found seven relevant papers (Guan et al., 2007), (Rigutini et al., 2005), (Guo and Xiao, 2012), (Cui et al., 2011), (Yu et al., 2010b), (Wu et al., 2019), (Jing et al., 2017)

The third plot from the left of figure 2 illustrates that for multidimensional inputs we found one paper of interest (Colares et al., 2013), for sequential inputs we found three papers of interest (Shi et al., 2010), (Huang and Hasegawa-Johnson, 2009), (Nigam et al., 2006) and for univariate inputs we found no paper of interest.

The fourth plot from the left of figure 2 Illustrates that for multidimensional inputs we found three papers of interest (Liu et al., 2016), (Rajan et al., 2008), (Minakawa et al., 2013) for sequential inputs we have found six relevant papers (Yu et al., 2010a), (Zhu et al., 2009), (Zhu et al., 2008), (Zhang and Chen, 2002), (Varadarajan et al., 2009), (Kim et al., 2006) and for univariate inputs we did not find any relevant papers.

From figure 1 and we can confirm that the most popular methods are based on co-training and multi-view learning, graph-based methods, mixture models and uncertainty sampling. Clearly semi-supervised methods are more popular than active learning methods, three to one. Uncertainty sampling however includes many ways to measure uncertainty so one could argue that we should divide it into subcategories.

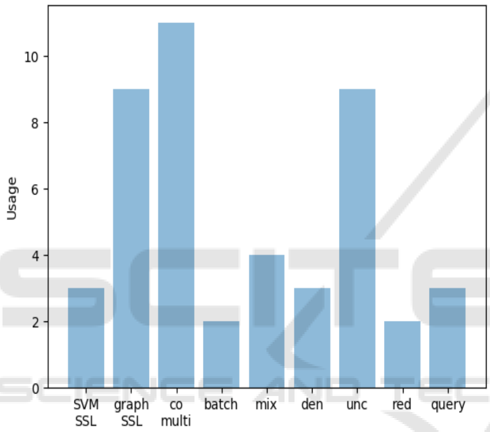

Figure 1: Overview showing the distribution of each method over the papers studied in this article.

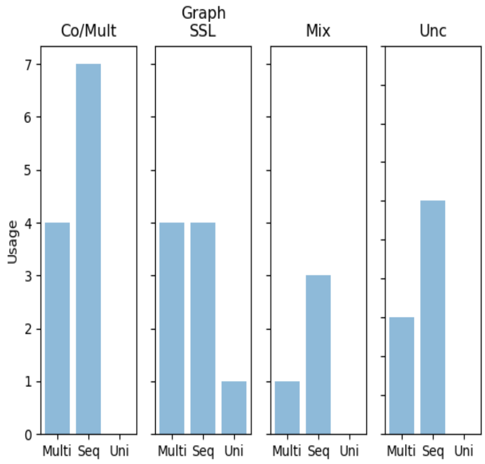

Figure 2: Overview showing the distribution of datatypes over methods based on uncertainty sampling.

6 DISCUSSION

In the research section we found what semi-supervised methods and active learning methods are popular. It is important to highlight that most of the papers from this study are not based entirely on the methods examined but are all some kind of hybrid with other types of learning.

In the background we presented several issues that concerned missing labels in company data and that it is expensive so fix this issue. From reviewing all the papers in this review, we found that the common factor they all shared was that they were all missing labels and they had a hard time obtaining these because of financial and labor costs.

None of the articles compared different semi-supervised learning algorithms with each other so we cannot compare the accuracy of each method. This is because each method has its own distinct assumption in order to work properly (Zhu and Goldberg, 2009). Therefor it is easy to predict that some methods with different assumptions will not work on the same data as others. Thus, a comparison of semi-supervised algorithms is not necessary.

Active learning methods was compared thoroughly in the papers. The best query strategy could not be identified but the empirical evidence suggests that every active learning approach exceeded the random sampling approach. This proves that active learning is much more effective than choosing the instances to be labeled randomly.

7 OPEN RESEARCH QUESTIONS

The semi-supervised methods described in this paper are the most basic ones and are taken from (Settles, 2012) and the active learning methods are taken from (Zhu and Goldberg, 2009). Most papers studied are based on active and semi-supervised learning algorithm. Some of the methods in the papers resembles these methods and some does mot, and this is why it is hard to compare every method to each other so there is a lot of research in just improving old methods.

There are many open research questions available, e.g as society is becoming more data-driven, we need to know how do to efficiently incorporate large or infinite amounts of data into our labeling algorithms.(Settles, 2012). Researchers wish to create semi-supervised learning algorithms that preform better than supervised learning by selecting the best semi-supervised parameters and assumptions for the model(Settles, 2012). Ideally semi-supervised learning should be used with all types data from different

Table 1: Categories for each application.

Category	Datatype	Area
Multidimensional inputs	Image, Video	Image classification/segmentation Image retrieval. Detection in videos Monocular 3D human pose estimation. Microalgae classification.
Sequential inputs	Time Series, Signals, Text.	Text classification, segmentation Word-sense disambiguation Signal processing Spoken language understanding/Speech recognition Word segmentation Phonetic classification Information extraction/retrieval
Univariate inputs	One-dimensional	Real time traffic classification Webpage classification Network intrusion detection

Table 2: Summary of all methods for ”active” and ”semi-supervised” learning.

Active Learning	Semi-supervised Learning
Uncertainty Sampling	Semi-supervised SVM
Query by Committee	Co-training
Query by Disagreement	Mixture Models
Expected Model Change	Cluster-then-label
Expected Error Reduction	EM algorithm
Density-Weighted methods	Multi-view learning
Variance Reduction	Graph-Based
Cluster Based	

Table 3: Classification of semi-supervised methods.

Category	Sub-ategory	Methods
Semi-supervised	co-multi	Co-training and multi-view learning
	mix	Mixture models.
	SVM	Support Vector Machines.
	graph	Graph Based method.
Active	den	Density weighted methods.
	query	Query-by methods e.g QBC and QBD.
	red	Expected error or variance reduction.
	unc	Uncertainty sampling based methods.

areas. To make semi-supervised work on all these different datatypes, we need to define new assumptions for the models and its parameters(Settles, 2012). An impressive field of study is combining active learning and semi-supervised learning. Active learning is first used to determine what instances to label. These manually labeled instances will then be used for the semi-supervised part of the model. (Weston et al., 2012) For more applications see (Hakkani-Tur et al., 2011), (Tur et al., 2005), (Zhu et al., 2003), (Leng et al., 2013).

In future research we would like to explore the following:

- How can we combine active learning with semi-supervised deep learning models? Semi-supervised learning relies heavily on data assumptions. Deep learning however does not rely on the structure of the data.
- How can we train automatic labeling algorithms with additional infrastructure e.g test lab equipment?
- How do we use time as a mechanism for automatic labeling? When predicting an outcome, how do we use the actual outcome that becomes available after some time.
- How sensitive are learning algorithms for noise and how low-quality data and what mitigation strategies exists?

8 CONCLUSION

Our goal of this study is to provide a structured overview over machine learning methods used for labeling unlabeled data and to identify the open research challenges associated with automatic labeling.

The basis of this problem comes from the industry rather than academia. Companies have a vast amount of data that is not useful for supervised learning tasks as these require labeled data. Since more than 95% of the deployments of artificial intelligence in industry, based on our observations, are concerned with supervised learning, having labels is crucial for companies and different strategies to obtain labels have been adopted. These include obtaining labels through crowdsourcing, hiring individual contractors or educate their own staff so that they can do the labeling manually. All of these approaches involve huge financial costs and laboring costs that the companies wish to reduce.

It proves to be difficult to find a fully automatic approach to labeling as most approaches needs human intervention of some sort. Human intervention in machine learning is discussed in interactive machine learning. Active learning is a brand in machine learning were we are allowed to be pose queries in order to choose what instances should be labeled to be included in the training set. Semi-supervised learning is a brand in machine learning where we use a small set of labeled instances to try and achieve better results that supervised learning algorithms.

Both active and semi-supervised machine learning algorithms can be used to solve problems in which we have an insufficient amount of labeled data, but they do this in different ways. Based on our analysis we can say that semi-supervised and active learning methods are well developed for labeling and between the two, semi-supervised learning seems to be more developed. Unlike active learning, semi-supervised learning does not require any human intervention and is therefore more ”automatic” and require less effort from humans. Furthermore, we see great potential in using the methods presented in this article for industrial applications and to contribute with new ideas especially to univariate data since the current research on this datatype is lacking. A particularly interesting research topic is to combine active learning with semi-supervised learning, e.g one could use active learning to pose queries in order to find the optimal instances to label for inclusion in the training data and then use semi-supervised learning for whatever purpose we want to use it for.

The contribution of this paper is threefold. First, we provide an overview of the available approaches for (semi-)automatic labeling of data for machine learning based on a systematic literature review. Second, we present the data types that are typically subject to (semi-)automatic labeling and which data types require additional research. Finally, we identify the open research questions that need to be addressed by the research community.

ACKNOWLEDGMENT

This work was partially supported by the Wallenberg AI Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Fundation.

REFERENCES

Abbasi, M., Rabiee, H. R., and Gagné, C. (2015). Monocular 3d human pose estimation with a semi-supervised graph-based method. In *2015 International Conference on 3D Vision*, pages 518–526. IEEE.

AzatiSoftware (2019). *AzatiSoftware Automated Data Labeling with Machine Learning*. https://azati.ai/automated-data-labeling-with-machine-learning.

Basu, S., Davidson, I., and Wagstaff, K. (2008). *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press.

Brefeld, U., Gärtner, T., Scheffer, T., and Wrobel, S. (2006). Efficient co-regularised least squares regression. In *Proceedings of the 23rd international conference on Machine learning*, pages 137–144.

Chen, C., Gong, Y., and Tian, Y. (2008). Semi-supervised learning methods for network intrusion detection. In *2008 IEEE international conference on systems, man and cybernetics*, pages 2603–2608. IEEE.

CloudFactory.com (2019). *The Ultimate Guide to Data Labeling for Machine Learning*. https://www.cloudfactory.com/data-labeling-guide.

Colares, R. G., Machado, P., de Faria, M., Detoni, A., Tavano, V., et al. (2013). Microalgae classification using semi-supervised and active learning based on gaussian mixture models. *Journal of the Brazilian Computer Society*, 19(4):411–422.

Cortes, C. and Mohri, M. (2007). On transductive regression. In *Advances in Neural Information Processing Systems*, pages 305–312.

Cui, X., Huang, J., and Chien, J.-T. (2011). Multi-view and multi-objective semi-supervised learning for large vocabulary continuous speech recognition. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4668–4671. IEEE.

Di, W. and Crawford, M. M. (2011). View generation for multiview maximum disagreement based active learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5):1942–1954.

Guan, D., Yuan, W., Lee, Y.-K., Gavrilov, A., and Lee, S. (2007). Activity recognition based on semi-supervised learning. In *13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2007)*, pages 469–475. IEEE.

Guo, Y. and Xiao, M. (2012). Cross language text classification via subspace co-regularized multi-view learning. *arXiv preprint arXiv:1206.6481*.

Hakkani-Tur, D. Z., Schapire, R. E., and Tur, G. (2011). Combining active and semi-supervised learning for spoken language understanding. US Patent 8,010,357.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Huang, J.-T. and Hasegawa-Johnson, M. (2009). On semi-supervised learning of gaussian mixture models for phonetic classification. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 75–83. Association for Computational Linguistics.

Jing, X.-Y., Wu, F., Dong, X., Shan, S., and Chen, S. (2017). Semi-supervised multi-view correlation feature learning with application to webpage classification. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Kääriäinen, M. (2005). Generalization error bounds using unlabeled data. In *International Conference on Computational Learning Theory*, pages 127–142. Springer.

Keele, S. et al. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, Ver. 2.3 EBSE Technical Report. EBSE.

Kim, S., Song, Y., Kim, K., Cha, J.-W., and Lee, G. G. (2006). Mmr-based active machine learning for bio named entity recognition. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 69–72.

Leng, Y., Xu, X., and Qi, G. (2013). Combining active learning and semi-supervised learning to construct svm classifier. *Knowledge-Based Systems*, 44:121–131.

Li, Y. and Guan, C. (2008). Joint feature re-extraction and classification using an iterative semi-supervised support vector machine algorithm. *Machine Learning*, 71(1):33–53.

Liu, P., Zhang, H., and Eom, K. B. (2016). Active deep learning for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2):712–724.

Liu, Q., Liao, X., and Carin, L. (2008). Semi-supervised multitask learning. In *Advances in Neural Information Processing Systems*, pages 937–944.

Liu, Y. and Kirchhoff, K. (2013). Graph-based semi-supervised learning for phone and segment classification. In *INTERSPEECH*, pages 1840–1843.

Madani, O., Pennock, D. M., and Flake, G. W. (2005). Co-validation: Using model disagreement on unlabeled data to validate classification algorithms. In *Advances in neural information processing systems*, pages 873–880.

Minakawa, M., Raytchev, B., Tamaki, T., and Kaneda, K. (2013). Image sequence recognition with active learning using uncertainty sampling. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Morsillo, N., Pal, C., and Nelson, R. (2009). Semi-supervised learning of visual classifiers from web images and text. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Nigam, K., McCallum, A., and Mitchell, T. (2006). Semi-supervised text classification using em. *Semi-Supervised Learning*, pages 33–56.

Rajan, S., Ghosh, J., and Crawford, M. M. (2008). An active learning approach to hyperspectral data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4):1231–1242.

Ranzato, M. and Szummer, M. (2008). Semi-supervised learning of compact document representations with deep networks. In *Proceedings of the 25th international conference on Machine learning*, pages 792–799.

Rigutini, L., Maggini, M., and Liu, B. (2005). An em based training algorithm for cross-language text categorization. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 529–535. IEEE.

Rosset, S., Zhu, J., Zou, H., and Hastie, T. J. (2005). A method for inferring label sampling mechanisms in semi-supervised learning. In *Advances in neural information processing systems*, pages 1161–1168.

Schuurmans, D. and Southey, F. (2002). Metric-based methods for adaptive model selection and regularization. *Machine Learning*, 48(1-3):51–84.

Settles, B. (2012). Active learning, volume 6 of synthesis lectures on artificial intelligence and machine learning. *Morgan & Claypool*.

Shi, L., Mihalcea, R., and Tian, M. (2010). Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067. Association for Computational Linguistics.

Sindhwani, V., Niyogi, P., and Belkin, M. (2005). A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, volume 2005, pages 74–79. Citeseer.

Stikic, M., Larlus, D., and Schiele, B. (2009). Multi-graph based semi-supervised learning for activity recognition. In *2009 International Symposium on Wearable Computers*, pages 85–92. IEEE.

Tang, J., Hong, R., Yan, S., Chua, T.-S., Qi, G.-J., and Jain, R. (2011). Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2):1–15.

Tang, J., Li, H., Qi, G.-J., and Chua, T.-S. (2008). Integrated graph-based semi-supervised multiple/single instance learning framework for image annotation. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 631–634.

Tang, J., Li, H., Qi, G.-J., and Chua, T.-S. (2009). Image annotation by graph-based inference with integrated multiple/single instance representations. *IEEE Transactions on Multimedia*, 12(2):131–141.

Tur, G., Hakkani-Tür, D., and Schapire, R. E. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186.

Varadarajan, B., Yu, D., Deng, L., and Acero, A. (2009). Maximizing global entropy reduction for active learning in speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4721–4724. IEEE.

Weston, J., Ratle, F., Mobahi, H., and Collobert, R. (2012). Deep learning via semi-supervised embedding. In *Neural networks: Tricks of the trade*, pages 639–655. Springer.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10.

Wu, F., Jing, X.-Y., Zhou, J., Ji, Y., Lan, C., Huang, Q., and Wang, R. (2019). Semi-supervised multi-view individual and sharable feature learning for webpage classification. In *The World Wide Web Conference*, pages 3349–3355.

Yan, R. and Naphade, M. (2005). Semi-supervised cross feature learning for semantic concept detection in videos. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 657–663. IEEE.

Yu, D., Varadarajan, B., Deng, L., and Acero, A. (2010a). Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Computer Speech & Language*, 24(3):433–444.

Yu, Z., Su, L., Li, L., Zhao, Q., Mao, C., and Guo, J. (2010b). Question classification based on co-training style semi-supervised learning. *Pattern Recognition Letters*, 31(13):1975–1980.

Zeng, X., Wong, D. F., Chao, L. S., and Trancoso, I. (2013). Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 770–779.

Zhang, C. and Chen, T. (2002). An active learning framework for content-based information retrieval. *IEEE transactions on multimedia*, 4(2):260–268.

Zhang, C. and Zheng, W.-S. (2017). Semi-supervised multiview discrete hashing for fast image search. *IEEE Transactions on Image Processing*, 26(6):2604–2617.

Zhao, M., Chow, T. W., Zhang, Z., and Li, B. (2015). Automatic image annotation via compact graph based semi-supervised learning. *Knowledge-Based Systems*, 76:148–165.

Zhou, Z.-H. and Li, M. (2005). Semi-supervised regression with co-training. In *IJCAI*, volume 5, pages 908–913.

Zhou, Z.-H. and Xu, J.-M. (2007). On the relation between multi-instance learning and semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1167–1174.

Zhu, J., Wang, H., Tsou, B. K., and Ma, M. (2009). Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6):1323–1331.

Zhu, J., Wang, H., Yao, T., and Tsou, B. K. (2008). Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144.

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

Zhu, X., Lafferty, J., and Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, volume 3.