Summary of Data Label Research under Smart Finance

Chao Yu[1], Zhiqiang Gao[2], Xiaolong Cui[2], Shaozhe Guo[2]
1838923947@qq.com, 1090398464@qq.com, ecxl@qq.com, 2190962168@qq.com

[1]Engineering University of PAP, Electronic and communication engineering, Xi'an, China

[2]Engineering University of PAP Xi'an, China

Abstract—How financial product managers better understand and manage users involves the use of data tags. From the perspective of label, this paper expounds the generation method of data labels, summarizes the classic common data set, open source data annotation tools and commercial crowdsourcing platform, expounds the different application scenarios of data labels under big data, and briefly summarizes the tag system and user portrait system. Finally, the development of labeling from labeling requirements, labeling efficiency and data security is discussed, and a distributed labeling system is proposed for data security problems.

Keywords-data label; artificial intelligence; big data; finance

1 INTRODUCTION

AI + finance refers to enabling all the participants and business links of the financial industry through artificial intelligence technology as the main driving force, highlighting the important role of AI technology in product innovation, flow reengineering and service upgrading of the financial industry. In recent years, data labeling, as the data support of artificial intelligence, has promoted the profound development of machine learning. In 2014, Alibaba and Ali Mom jointly hosted a competition called -- Data Scientists Competition, a big data commercialization competition based on a real business environment. According to the marketing needs of merchants and with the help of Alibaba's open data computing power, the participants will depict the consumer group, conduct data algorithm mining, conduct marketing scheme design, and realize accurate delivery. First of all, we should deeply understand users and identify users, including the basic information of borrowers, financing preferences and credit investigation conditions. To do this well, it is inseparable from the user data label construction. After labeling user data, user data is analyzed through machine learning methods. This paper describes the basic role and application of labels in big data analysis through the generation, application and quality analysis, and then provides services for financial data analysis.

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328436

2 DATA ANNOTATION GENERATE LABELS

2.1 Introduction of data annotation

Data annotation provides resource support for machine learning, its specific operation is: (1) label part of the data that needs to be learned. (2) Build an algorithmic model. And optimize the model data through continuous iterations. (3) test the model by using new, untrained data. after learning a certain amount of data, make the machine to realize independent identification characteristics, can carry out automatic annotation, classification and other operations. This is equivalent to the process that human beings acquire new knowledge and can apply it with a certain accuracy. The development of data annotation has promoted the vigorous development of artificial intelligence, and machine learning supported by data annotation is widely used in various fields[1]. With the rapid development of artificial intelligence technologies such as machine learning and computer vision, the development of smart finance has gradually become mature. It is mainly used in eight scenarios, including intelligent risk control, intelligent insurance, intelligent customer service, intelligent marketing, intelligent supervision, intelligent investment consulting, intelligent investment and research, and identity identification.

- Data annotation object

Data-labeled objects include text, voice, pictures, and videos. Different data require different annotation methods. Cai Li, Wang Shuting, et al[1]listed the annotation methods of the different data in the Survey of data annotation and further refined for the different annotation tasks.

Table 1 Classification of data annotation [1]

Classification method	classification method	concept	merit	shortcoming
Dimension object	Image annotation	Image annotation and video annotation are collectively known as image annotation	Face recognition and autonomous driving technologies have been developed and improved	Relatively complex, and time-consuming
	Voice annotation	Language content needs to be manually transcribed into text content, and then identified by the algorithmic model	Help to more improve the voice recognition function in the field of artificial intelligence	The speech content cannot be directly understood, and text transcription is required
	Text annotation	Some similar to audio annotation, they need to be transcribed into text form	Reduced manual workload in the text recognition industry and field	The manual identification process is complicated
The composition form of the dimension	Structured annotation	The data label must be within the specified set of label candidates. The annotator selects the most reasonable label	The label candidate set describes label categories clearly for them to choose; labels are structured for storage and later statistical search	Having ambiguous labels can often affect the final annotation results

		value as the annotation result by matching the label object with the set of label candidates		
	Unstructured annotation	The free keyword describes the labeled object within the specified constraints	Give the label name enough freedom to clearly express your opinion	It brings difficulties to data storage and use, and is unfavorable to statistical analysis
	Semi-structured annotation	Label values are structured labels, while the label domain is unstructured with labels	Lanotation flexibility, easy to find statistics	High labeling requirements, high workload, and time-consuming
Types of annotators	Manual annotation	Emploire trained annoator to mark	High labeling quality	The labeling cost is high, long time and low efficiency
	Machine annotation	The annotator is usually an intelligent algorithm	The labeling is fast and at relatively low cost	The algorithm performs poor recognition and extraction of objects involving high-level semantics

Table 1 details the classification of data annotation from the three aspects of annotation objects, annotation composition format, and annotation person type, and summarizes its corresponding advantages and disadvantages. For the annotation content, simple text, picture annotation, we only need to deal with the one dimension annotation problem. For the video, audio and other data, or the text recognition in the picture, we need to process the multi-dimensional information and annotate it at the same time. The annotation format is also multi-modal. During the annotation of the data, Gong Chen et al[18] propose that the bias of class-related labels is mainly affected by the deviation of subjective cognition, so that different annotators may give different labels. At the same time, the data labels have a certain correlation[19], By analyzing their associations and differences, we can find the information that we need.

- Data annotation mode

Common data annotation modes include classification annotation, frame annotation, area annotation, punctuation annotation, etc.: among them, frame annotation is mainly used for machine vision, and the objects in the picture are marked. widely used in face recognition, item identification, etc. Area annotation refers to the labeling mode with the edge of flexible lines, which is more accurate and therefore can be used for driverless driving. punctuation marking is often applied to tasks that require detailed eigenvalues, such as face recognition. There are also other markers, such as semantic annotation, voice annotation, etc. Figure 1~4 shows some instances of data annotation:

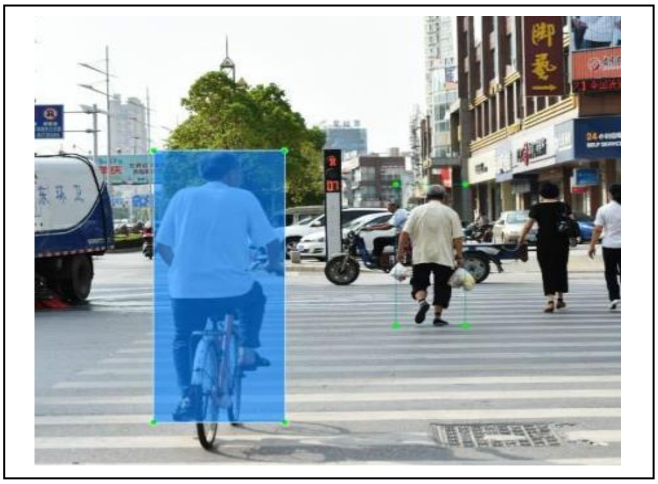


Figure 1. box Select dimension.

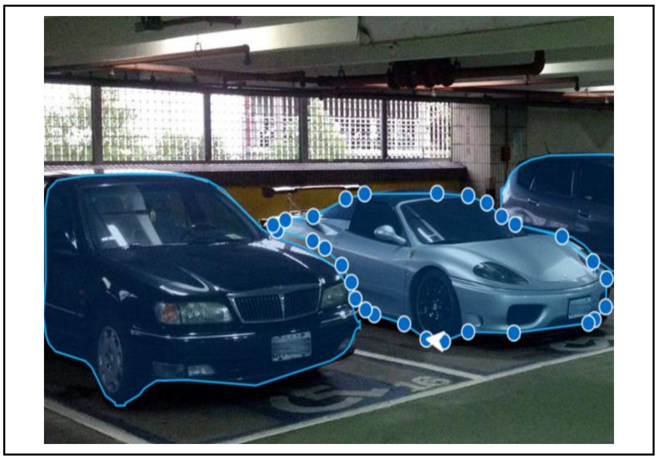


Figure 2. area dimension.

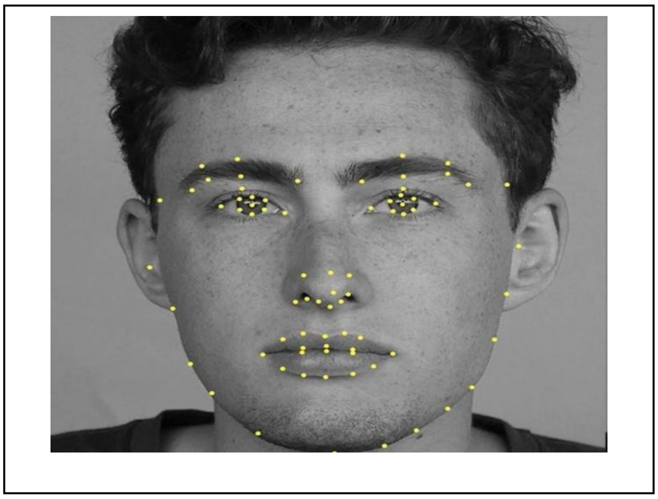
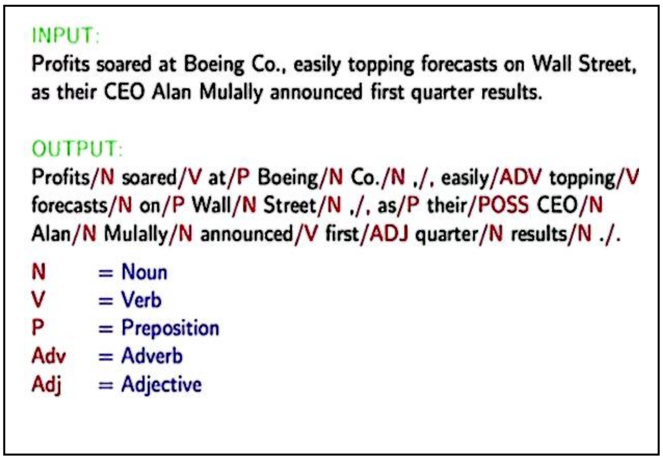


Figure 3. Punctuation annotation.



INPUT:
Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:
Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

N	= Noun
V	= Verb
P	= Preposition
Adv	= Adverb
Adj	= Adjective

Figure 4. Semantic annotation.

In terms of the data format that needs to be annotated, the annotation can be divided into the annotation of text, audio, image and video. Text annotation mainly uses labels to highlight keywords, phrases, and sentences to make the machine correctly recognize the human language. Including emotion, intention, semantics, named entity annotation, etc. widely used in various financial institutions, provide 24-hour question and answer service; in technology, it is based on natural language processing, speech recognition and other technologies to create intelligent question and answer system. The annotation of speech includes the transcription of specific language and intonation, feature recognition, etc. Play a key role in customer service and telemarketing. Image annotation typically provides solutions for computer vision, machine recognition, and face recognition. Video annotation is usually to annotate the video keyframes or time to achieve the purpose of identification and tracking. In the financial industry is mainly used in face recognition, customer interaction and other service fields. Financial time series can also do the corresponding feature extraction, such as yield, volatility, various technical indicators: moving average, momentum, etc., of course, can also be financial characteristics such as PE, PB, ROA, etc. The characteristics here can correspond to the "factors" in the traditional financial analysis. We can use the machine learning model to find the data association based on the traditional alpha factors.

2.2 Existing datasets and data annotation methods

Currently, there are thousands of open-source machine learning datasets on the network, each of which corresponds to different annotation tasks and can achieve different functions. At the same time, for a new task, different data types and task needs require people to use different annotation methods. This section give some lists about the open-source commonly used data annotation data sets, data annotation tools, and commercial annotation platforms.

- Data annotation dataset

Table 2 lists some commonly used open-source data annotation datasets, and gives the specific role, size, and download address of the datasets accordingly. Among them, the COCO (Common Objects in Context) dataset sponsored by Microsoft is an image recognition, segmentation and image semantic data set. Its image annotation information has not only category and location information, but also the semantic text description of the image, which has become the "standard" data set for the performance evaluation of the image semantic understanding algorithm. The ImageNet is an image dataset organized according to the WordNet hierarchy. The WordNet contains approximately 100,000 words, and the ImageNet provides on average about 1,000 images to illustrate each word. The Yelp dataset has 5,200,000 reviews, 174,000 commercial attributes, 200,000 photos, etc. The Million Songs dataset is a million contemporary pop music repertoire of audio features and metadata, at the core of the feature analysis and metadata of a million songs. The dataset contains no audio and contains only derived features. In addition, there are datasets in the fields of face recognition, disease treatment, and financial services.

Table 2 Some commonly used annotated datasets

CLASS	**NAME**	**USE**	**BIG OR SMALL**	**SOURCE**	**WHETHER OPEN SOURCE**
PICTURE	MNIST	hand-written numeral recognition	50MB	https://datahack.analyticsvidhya.com/	yes
	MS-COCO	Objects detection and identification	40GB	https://cocodataset.org/	yes
	ImageNet	image classification object detection	150GB	https://image-net.org/	yes
	Open Images	image classification Positioning, detection	500GB	https://github.com/openimages/	yes
TEXT	Sentiment 140	sentiment analysis	80MB	http://help.sentiment140.com/	yes
	WordNet	semantic association	10MB	https://wordnet.princeton.edu/	yes

	Yelp	sentiment analysis	2.66 GB JSON 2.9 GB SQL A 7.5 GB picture	https://www.yelp.com/	yes
AUDIO FREQUENCY	FMA	Music type analysis	1000GB	https://github.com/	yes
	Ballroom	music style	14GB	http://mtg.upf.edu/	yes
	LibriSpeech	Language model training	60GB	http://www.openslr.org/12/	yes
VIDEO FREQUENCY	ALOV300++	Object tracking	10GB	http://alov300pp.joomlafree.it/	yes
	HMDB	movement identification	5GB	http://jhmdb.is.tue.mpg.de/dataset	yes

- Data annotation tool

Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".

Table 3 The Data Dimension Tool

NAME	NOTE THE FORM	FUNCTIONAL SUPPORT	ENVIRONMENTAL FOUNDATION	CHARACTERISTIC	EXPORT FORMAT
LABELLMG	Box selection	image detection	Python&Ixml	QT graphical interfaces	XML form
LABELLME	Polygon, drawing lines, Scaling, circular	Image classification, semantic segmentation, and video annotation	Python QT	Can video annotation	Json form
RECTLABEL	Polygon, drawing lines, Scaling, circular	object detection image segmentation	\	\	Coco, csv format
OPENCV/CVAT		Image classification, object detection, and semantic segmentation	Locally deployed	Has an integrated semi-automatic model	
VOTT	Polygons, punctuation	Image and video data annotation	JavaScript	WEB way	Vott form
LABELBOX	Polygon, drawing lines, punctuation	object detection, instance segmentation	Pure JS + HTML operation support	WEB mode custom annotation API support	Json form
VIA	Polygon, drawing lines, Scaling, circular	Target detection, image segmentation, and instance segmentation	Locally deployed	WEB mode, to provide face marking	Json form
POINT-CLOUD-ANNOTATION-TOOL	KITTI-bin1111111	3D BOX generation	Point cloud data loading	The 3D point-cloud data annotation work	txt document

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic.

- Data annotation platform

Due to the development of data magnitude and the demand of the algorithm, it is unlikely to obtain its own data set through its own annotation. For example, a data set containing 200,000 pictures. Assuming that the efficiency of a person is 2,000 per day, it takes 100 days to label it, which obviously does not meet the efficiency requirements. Therefore, a large number of data annotation tasks are usually delivered to the data crowdsourcing platform, which completes a large number of data annotation tasks, or seeks for some function expansion and application research[36]. At present, the largest crowdsourcing platforms in China include JD Zhongzhi Platform, Baidu crowdsourcing, Alibaba crowdsourcing, iFlytek, Totoro crowdsourcing, Haitian Ruisheng, Data Hall, Digital Plus, and Desai, etc. Among them, JD Zhongzhi has a fast feedback speed and can issue a data isolation scheme to ensure user data security; Baidu crowdsourcing has a wide business scope and provides perfect services. Totoro crowdsourcing, Haitian Ruisheng, Data Hall, Digital Canada, Canada and Desai and other companies specialize in data services, and the data quality is stable.

3 LABEL APPLICATIONS BASED ON DEEP LEARNING ALGORITHMS

After labeling large amounts of data, we obtained datasets with data labels.Now research usually combines labeled datasets with different machine learning algorithm algorithms to achieve different functional requirements.Based on the existing research results, there are the following categories: combined with recommendation algorithm, label recommendation, personalized recommendation, combined with communication algorithm to discover unknown network and analyze network characteristic performance, combined with clustering algorithm to improve label application quality; weight algorithm and classification algorithm, analyze the importance of labels to different objects, so as to form accurate data portrait and realize accurate marketing.

3.1 Label-recommendation algorithm

Pengyu Xu, Huafeng Liu ,et al[38]summarized the method of label recommendation in figure 5, proposed the method to assist users in data labeling through the label recommendation method, and improve the quality of label generation by recommending high-quality annotations.

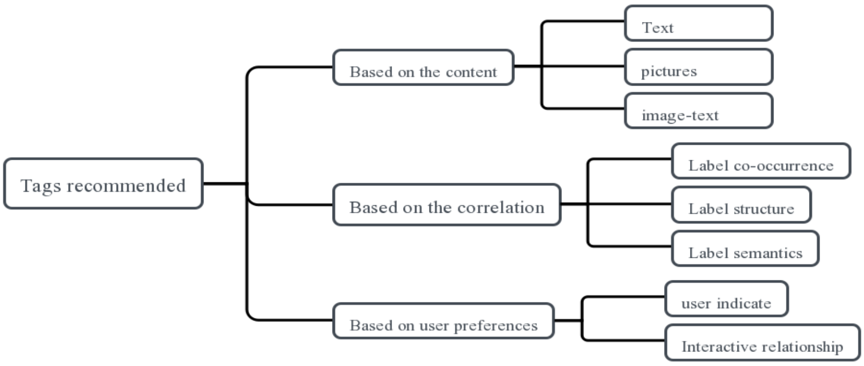


Figure 5. Recommended labeling method

Chen Hailong and Yan Wuyue[2]Improve the quality of interest recommendations by using collaborative filtering methods. Li Hongmei, Diao Xingchun, etc[9]propose a personalized label recommendation model, and explain the sparsity of the label data. Potential label mining and fine-grained preference relationships to form label recommendations based on Lastfm and Movielens datasets. Through label recommendation, appropriate label names can be provided for label personnel in the process of label production, so as to promote the improvement of labeling quality. Secondly, label recommendation can provide users with interest prediction in combination with collaborative filtering and weight algorithms.

3.2 Label propagation algorithm

The Label propagation (LPA) algorithm formed by the combination of data label and data propagation can be well applied in data mining, network rules, performance analysis, community discovery and other fields. Chen Hongtao, et al[3]summarized and analyzed the existing label propagation algorithm, and proposed a modified label propagation algorithm (Link Label Propagation Algorithm, LLPA).Geng Luyang, et al[5]propose a label propagation algorithm for adaptive control distance on account of the SLPA algorithm and the decay label propagation algorithm. Liu Shichao et al[13]proposed overlapping community discovery algorithm based on the label propagation probability(Label-Propagation-Probability-Based, LPPB), in addition, based on the clustering method , Wang Xili, et al [29] propose a semi-supervised propagation algorithm for finding a minimal cost path, Solve the problem of label classification with a large amount of data. The core idea of label propagation algorithm is very simple, that is, the same data should have the same label. Because the algorithm is simple and easy to implement, the algorithm execution time is short, low complexity and good classification effect, it has attracted the attention of scholars at home and abroad, and it is widely applied to multimedia information classification, virtual community mining and other fields.

3.3 Multi-label classification learning algorithm

Behaviors in social networks are massive and diverse. We want to predict the behavior in the network. This problem can be translated into a multi-label classification problem, and many scholars have achieved many results by combining labels with machine learning methods. Lin Xiao , Boli Chen [34] proposed the LASA (multi-label text classification based on label semantic attention) algorithm to obtain different text weights, consider label relevance, and achieve good classification effect. Qingyao Wu[20]proposed a novel algorithm: ML-FOREST, reveal intrinsic label dependencies by learning the ensemble of a hierarchical multi-label classifier tree . Fu zhongliang, et al[4] proposed a multi-label cost-minimized cost classification integrated learning algorithm. Compared with the multi-classification cost-sensitive algorithm, it has better results in solving the multi-class cost-sensitive classification problem. Danpu Zhang et al[43]proposed a label matching integrated learning algorithm to solve the dual-label set learning problem, but there are problems with special application scope and neglect of label relevance. While the study proposed by Sujuan Hou[21], Xiaoya Che[35] considers the label correlation and proposed their own approach for multi-label learning. But in reality, the data tags may be missing, so Yue Zhu, James T.Kwo ket al[41] propose a multi-label learning methods based on global and local label correlations, by learning potential label representations and optimizing label manifolds and exploiting both global and local label correlations, training classifiers and recovering missing labels.

3.4 Label-clustering algorithm

Classification and clustering methods seem to be the same, but in fact, classification refers to the process of mining data from specific data and making judgments, while clustering is to study the characteristics of different types of data labels and form a set of discriminative patterns. Clustering can therefore help to improve the algorithmic efficacy. The current clustering methods combined with labels are mainly divided into K-Means clustering, and the clustering methods based on the condensate hierarchy are less studied. Wang Feng and Hou Yanwei[24]et al combined the K-Means algorithm with the graph-based label propagation algorithm, first using the clustering algorithm to find the potential distribution structure in the data, and then performed label propagation in the cluster division obtained from the clustering to improve the quality of label propagation. Wen Liuying and Yuan Wei[32]Improving the effect of multi-label classification with K-Means clustering, propose the multi-label symbol value division and achieve the effect of data compression. It has a good guiding significance for the guiding attribute classification problem. Wang Hailong, Liulin[25] built a combination of MIR and k-means label clustering personalized recommendation algorithm in order to get better recommendation quality, form multi-modal network on the basis of the choice relationship between users and resources. Using K-means label cluster to search neighborhood users obtain the nearest neighbor user set, set the initial clustering center, according to the sorted recommended resources prediction results, realize personalized recommendation.

3.5 Label weighting algorithm

Kong Xinxin[8] put forward a label recommendation model based on the label weight score, used the weights to obtain the most accurate evaluation and needs of users, and solved the label recommendation problems such as cold start, complex interest recommendation and poor interpretability.

User portrait is a kind of data visualization technology generated to cope with too large data scale. It can extract and transform data and build data portrait function and system according to different needs[30]. Wu Jianyun, Xu Mingzhu[33],Based on the analysis of user purpose, others form a single user portrait based on user interests, and explore group user characteristics through clustering and theme, introduce time-index attenuation user labels, combined with affection and collaborative filtering for video recommendation ; Rui Zhou COVID-19 combined with big data portrait, described the application in public health events[47]; Under the concept of smart city, Ma Chao et al[15]proposed the portrait construction based on big data under the urban three-way space.

4 DATA LABELING SYSTEM AND QUALITY CONTROL

4.1 Data Label system

At present, many large enterprises and departments, based on the research on large data labels, study their own task needs, establish data platforms, and establish their own label system on the platform, so as to achieve big data-related business services. For example, Wang Xia and Xu Xiangdong[28]gave the Research on the construction method of health care big data labeling system . The data label system of public security system[44,46,48],And some comprehensive label system studies[42].Based on the labeling system, the knowledge map can be further constructed[40], using labels to evolve supernetworks [26,27];And apply the data label build and manage the data platform[12]. Figure 6 briefly summarizes the overall framework of the label system. The data labeling system is described from the generation framework of data labels, the use of labels, the management framework, and the system framework.

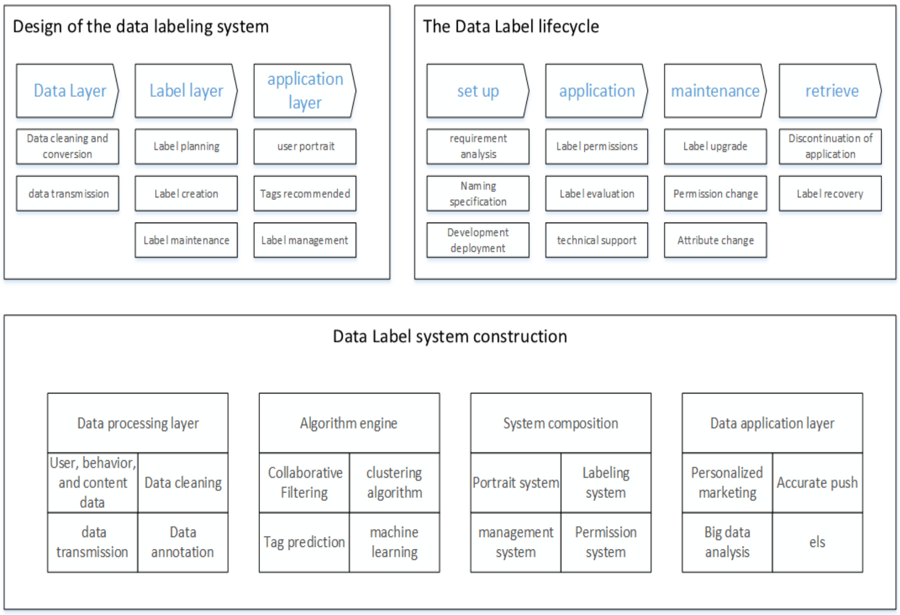


Figure 6. Composition of the data label system

4.2 Data Portrait system

Wang Dongyu[23] built the portrait on the basis of behavior analysis, and divided the user portrait into a static and dynamic mode to build the portrait system.

For user portrait of user label library function architecture, in which the data base management function layer contains the user portrait of data storage, system metadata management, label life cycle management, user portrait system query mechanism and update mechanism, the model label layer describes the static, dynamic information, data layer describes the data basic operation. Li Wangyue et al[10]. explored three sets of big data calculation methods and visualization methods on the basis of three labeling methods of original indicators, knowledge map and policy text; In addition, the user portrait also has a construction method based on text mining[14], Some communication companies use big data (network crawler, distributed storage technology, etc.) to obtain user information and annotate them, and using it realize the construction of the portrait system[31]. Generally, the user's basic attribute information is well stable, but the user's preferences and actions will change with time. At the same time, due to the emergence of new things will lead to the occurrence of new behaviors, the user preferences will also change, and the user behavior data will continue to change. Old data will lose value; Second, design user portraits in addition to basic information. In a word, the user portrait construction mainly considers the comprehensive design from the data type, user group, application scene (mode), implementation mode and other aspects, so as to scientifically manage and visualize the big data, so as to achieve the purpose of solving the user problems and realizing the operation functions.

4.3 Quality control

In the process of data labeling and label generation use, how to determine its application efficiency is also a big problem. Different tasks will produce different data, and the focus of their data labels and the quality evaluation criteria will be different, so different quality assessment methods will need to be adopted according to the task needs.

For the image annotation task, the main voting method is Majority Voting (MV)[22], Dawid & SKene (D & S) algorithm, expectation Maximization (EM) algorithm [37][39], RY truth inference algorithm, label aggregation algorithm KOS and other evaluation methods.

For speech annotation, there are quality studies on the complex correlation and similarity of semantics[6]. For the multimodal content, there will be different types of elements in the quality evaluation, and it is more difficult to formulate the quality evaluation standard.

Table 4 notes the quality evaluation algorithm [1]

classify	Algorithm	merit	shortcoming
Evaluation algorithm of image annotation quality	MV	easy to use, often used to do other crowdsourcing quality: the benchmark algorithm of the control algorithm	The different reliability of each annotation task, annoator was not considered
	EM	In a certain sense, it can converge to the local maximization	The convergence rate is relatively slow when the proportion of data missing is large

	RY	The classifier was combined with the Ground-truth for learning	Priors need to be imposed on the specificity and sensitivity of the annotation expert
Evaluation algorithm for the quality of text annotation	BLEU	Convenient, fast, and the results have a reference value	The evaluation accuracy is easily disturbed by common words
	ROUGE	The more reference annotations, the higher the correlation of the data to be evaluated	Unable not evaluate the fluency of the annotation data
	METEOR	Synonymous matching was considered for the evaluation, improving the evaluation accuracy	Length penalty, high measurement accuracy when the assessed data reclamation is small
	SPICE	Image annotation was evaluated from the semantic level of the graph	The semantic resolution of the graph needs to be further improved
	ZcnCrowd	Combining algorithm matching and manual matching realize the common improvement of annotation quality and efficiency to some extent	You cannot automatically select the best dataset for the fixed entity
Evaluation algorithm of speech annotation quality	WER	You can be divided into numbers, English, Chinese and other situations respectively	The effect is particularly bad when the data volume is large
	SER	The overall evaluation of sentences is better than the WER algorithm	Sentence error rate is higher, generally 2-3 times the word error rate

The table combines the different quality evaluation algorithms according to the different data formats, and lists the advantages and disadvantages. However, because many of the current data annotation work relies on crowdsourcing. Therefore, many scholars have put forward the quality evaluation and control methods based on the crowdsourcing methods. To solve the corpus quality problem in machine learning, Nie Zhenyun[17] proposed a crowdsourcing data annotation system based on personalized recommendation. Based on the EM algorithm, Xu Meng[38] propose crowdsourcing quality control methods based on worker ability and classification, increase the incentive mechanism and filter out low-quality and malicious workers, and improve the quality of annotation tasks. Yang Peijun[39]improved the effect of ELICE algorithm and EM algorithm by estimating the annotation ability and infer the label true value. In addition, there are also quality evaluation and improvement methods for crowdsourcing data[39][11][45]. Ipeirotis P G, Provost F[7]proposed A repeated-labeling strategy, Reduce label noise while using methods that improve data quality. Rafael Muller, Simon Kornblith, et al[16]Using label smoothing to neural network training, improved the training quality of labels and proposed a new research direction: the relationship between label smoothing and information bottleneck principle, with implications for data compression, generalization and information transmission in label use.

5 DIFFICULTIES AND OUTLOOK

As the basis of the application of artificial intelligence technology in the big data platform, the importance of data labeling is self-evident. In the field of e-commerce, everyone sees different products, and the system recommended to you is different; at the same time, everyone sees different information, the system will constantly recommend similar information to you according to your preferences. The system can combine the user according to the data and the third-party data, and then match the most appropriate push through the tag. However, from data labeling to labeling application, its development process still highlights some problems:

Problem 1: Requirements of different tasks: With the continuous development of deep learning and reinforcement learning, the quality requirements for labeling tasks are constantly increasing. Sometimes, it is necessary to provide new labeling methods for different financial needs to realize different functions. Therefore, the annotation personnel and annotation tools need to be constantly changed and upgraded.

Problem 2: annotation efficiency: At present, the domestic data annotation task demand is large, and it is inevitable for personnel to have low labeling quality when facing large tasks. However, the semi-automatic annotation technology through machine learning method is not mature enough. How to train a good model with a small number of manual annotation to achieve automatic annotation is still a big problem.

Question 3: Security issues: When using crowdsourcing-based annotation methods, data is at risk of leakage. For example, the labeling of identification information, some labeling related to business secret tasks, etc. At the same time, some enterprise platforms may transmit data to the Internet, individual users or to third-party platforms when completing the annotation task. Therefore, there is a risk of data leakage in the crowdsourcing annotation task.

For security problems, this paper proposes a distributed data annotation system in figure 7, which makes data dedense segmentation in different fields and formats of a task to ensure the low value and unresilience of segmented data secrets. The separated data will be transmitted to the corresponding professional department for annotation. After the annotation of each department, the data will be uploaded to the data center, and then the data will be annotated and stitched together to form a complete data annotation set. At the same time, set the label processing authority of different departments, and prohibit the unauthorized annotation persons from viewing the information outside the permission. And to ensure the time and space isolation of each department.

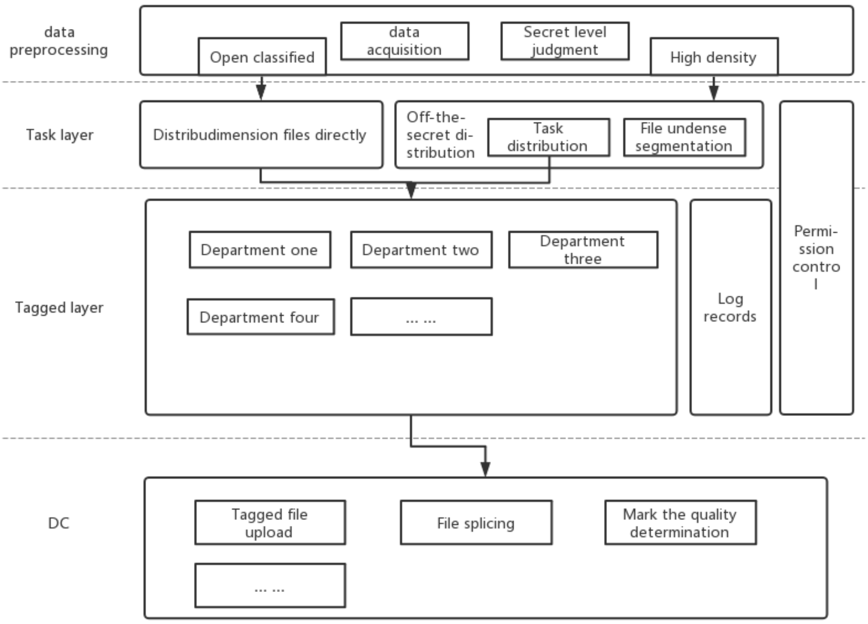


Figure 7. Distributed Data Dimnotation System

6 SUMMARY

Under the background of smart finance, take big data application and machine learning, summarize the generation methods of common data sets, open source data annotation tools and commercial crowdsourcing platform, combine different artificial intelligence algorithms, describe application scenarios of data tags, tag system and user portrait system, and propose a distributed annotation system for security issues. Finally, the development of labeling in the financial field is expected from the labeling demand, labeling efficiency and data security.

REFERENCES

[1] Cai Li, Wang Shuting, et al.Review of the data annotation studies.Journal of Software(in China), 2020,31 (2): 302320.

[2] Chen Hailong, Yan Wuyue, Sun Haijiao, Cheng Miao.Co-filtering recommendation algorithm for label extension [J].Computer Engineering and Science(in China), 2021,43 (10): 1826-1832.

[3] Chen Hongtao.Research and application of community discovery based on an improved label propagation algorithm [D].Hunan University(in China), 2018.

[4] Fu Zhongliang.Multilabel cost-sensitive classification ensemble learning algorithm.Journal of Automatics(in China), 2014,40 (6): 10751085

[5] Geng Luyang.Adaptive label propagation algorithm research and Application implementation [D].Chongqing University(in China), 2016.

[6] Guo Xiaoming, Ma Liangyu, Su Kai, Sun Yufei.Study on the automatic evaluation method of metadata quality based on semantic annotation [J].Computer Application and software(in China), 2018,35 (06): 23-27 + 82.

[7] Ipeirotis P G, Provost F, Sheng V S, et al.Repeated labeling using multiple noisy labelers[J].Data Mining and Knowledge Discovery, 2014, 28(2): 402-441.

[8] Kong Xinxin, Su Benchang, Wang Hongzhi, Gao Hong, Li Jianzhong.Recommended model and algorithm study based on label weight score [J].Journal of Computer(in China), 2017,40 (06): 1440-1452

[9] Li Hongmei, Diao Xingchun, et al. Personalized label recommendations based on potential label mining and fine-grained preferences [J / OL]. Computer Application Research (in China). https://doi.org/10.19734/j.issn.1001-3695.2018.05.0498.

[10] Li Wangyue, Liu Jin, Chen Na.The Application of Big Data Technology in Rural Portrait [J].Big Data(in China), 2020,6 (01): 99-118.

[11] Li Yinan, Wang Shitong.Methods for crowdsourced data [J].Journal of Intelligent Systems(in China), DOI: 10.11992/tis.201810014.

[12] Liu Peng.Design and implementation of the Spark-based data management platform [D].Hangzhou: Zhejiang University(in China), 2016.

[13] Liu Shichao, Zhu Husi, Gan Lin.Overlap community discovery algorithm based on label propagation probability [J].Journal of Computer Science(in China), 2016,39 (04): 717-729.

[14] Liu Xingchen.The Design and Implementation of a User Portrait System based on Text Mining [D].Beijing University of Posts and Telecommunications(in China), 2018.

[15] Ma Chao, Li Gang.Construction of urban data portrait based on urban big data [J].Modern Intelligence(in China), 2019,39 (08): 3-9.

[16] Müller R, Kornblith S, Hinton G.When does label smoothing help?[J].arXiv preprint arXiv:1906.02629, 2019.

[17] Nie Zhenyun.Crowdsourced-based data annotation system [D].Beijing Jiao Tong University(in China), 2014

[18] Gong Chen, Zhang Chuang, Wang Qizhou. Review of robust learning algorithms for label noise [J].Aviation weapons(in China), 2020,27 (03): 20-26

[19] Pan Chan, Feng Lifei et al.Label-keyword-based user behavior analysis [J].Intelligence Journal(in China), 2010,29 (03): 139-142.

[20] Qingyao Wu, Mingkui Tan,ML-FOREST: A Multi-label Tree Ensemble Method for Multi-Label Classification,IEEE Transactions on Knowledge and Data Engineering, DOI 10.1109/TKDE.2016.2581161.

[21] Sujuan Hou, Shangbo Zhou,Multi-label learning with label relevance in advertising video, Neurocomputing 171(2016) 932-848,Doi: http://dx.doi.org/10.1016/j.neucom.2015.07.022

[22] Tao F, Jiang L, Li C.Label similarity-based weighted soft majority voting and pairing for crowdsourcing[J].Knowledge & Information Systems(in China), 2020, 62(7).

[23] Wang Dongyu.The Design of a User Portrait System Based on Mobile Internet Behavior Analysis [D].Chengdu University of Technology(in China), 2017.

[24] Wang Feng, Hou Yanwei, Wei Wei, Cui Junbiao. Clustering-based label ensemble propagation algorithm [J / OL]. Journal of Shanxi University (Natural Science Edition)(in China) https: / / doi.org/10.13451/j.sxu.ns.2021002.

[25] Wang Hailong, et al. " Music personalized recommendation algorithm based on information retrieval and k-average clustering. Journal of Jilin University (Engineering edition) (in China)51.5 (2021): 1845-1850.

[26] Wang Jin, Liu Bin, Sun Kewei, Chen Qiaosong, Deng Xin. Multi-label evolution supernetwork based on label correlation [J].Electronic Journal(in China), 2018,46 (04): 1012-1018.

[27] Wang Wei ZhiHua Zhou, Crowdsourcing label quality: a theoretical analysis, Information Sciences, doi: 10.1007/s11432-015-5391-x.

[28] Wang Xia, Xu Xiangdong, Zhou Guanghua, Yang Zhe, Zhang Yuxi.Research on the construction method of healthcare big data label system [J].Chinese Journal of Health Information Management(in China), 2021,18 (02): 189-193.

[29] Wang Xili, Lin Hongshuai. Minimum-cost path label propagation algorithm [J].Journal of Computer(in China), 2016,39 (07): 1407-1418.

[30] Wang Yang, Ding Zhigang, Zheng Shuquan, Qi Wenxiu.The Design and Implementation of a User Portrait System [J].Computer Application and Software(in China), 2018,35 (03): 8-14.

[31] Wang Yilin.Implementation and Application of the Telecom User Portrait System Based on Big Data Technology [D].Beijing University of Posts and Telecommunications(in China), 2018.

[32] Wen Liuying, Yuan Wei.Cluster method of multi-label symbolic attribute values [J].Journal of Shandong University (Science edition)(in China), 2020,55 (03): 58-69.

[33] Wu Jianyun, Xu Mingzhu. Personalized recommendation based on user portrait and video interest tags [J].Intelligence Science(in China), 2021,39(01):128-134.DOI:10.13833/j.issn.1007-7634.2021.01.017.

[34] Xiao Lin, Chen Boli, Huang Xin, Liu Huafeng, Jing Liping, Yu Jian.Multi-label text classification based on label semantic attention [J].Journal of the Software Science(in China), 2020,31(04):1079-1089.DOI:10.13328/j.cnki.jos.005923

[35] Xiaoya Che, Degang Chen, A novel approach for learning label correlation with application to feature selection of multi-label data ,Information Sciences(2019), doi:https://doi.org/10.1016/j.ins.2019.10.022

[36] Xu Maocheng. The Design and Implementation of Baidu Mass Measurement and Mark System and its Extension Application in Data Acquisition [D].Nanjing University(in China), 2018.

[37] Xu Meng. Crowdsourcing quality evaluation and control method for the test question annotation task [D].Dalian University of Technology(in China), 2020.DOI:10.26991/d.cnki.gdllu.2020.002020.

[38] Xu Pengyu, Liu Huafeng, et al.Review of label recommendation methods studies.Journal of Software Science(in China), 2022,33 (4).http://www.jos.org.cn/ 1000-9825/6481.htm.

[39] Yang Peijun.An algorithm study for improving the annotation quality of crowdsourcing data [D].East China Normal University(in China), 2019.

[40] Yang Yuji, Xu Bin, Hu Jiawei, Tong Meihan, Zhang Peng, Zheng Li.An accurate and efficient method for constructing domain knowledge mapping.Journal of Software(in China), 2018,29 (10): 29312947. http: / / www.jos.org.cn/1000-9825/5552.htm

[41] Yue Zhu, James T.Kwok , Multi-Label Learning with Global and Local Label Correlation , IEEE Transactions on Knowledge and Data Engineering ,DOI :10.1109/TKDE.2017.2785795

[42] Zhan Qing.Design of a Big Data Parallel Classification System for Electronic Archive Sequences Based on Digital Label [J].Modern Electronics Technology(in China), 2020,43 (14): 152-155.

[43] Zhang Danpu, Wang Lili, Fu Zhongliang, Li Xin.Label matching ensemble learning algorithm based on two-label sets [J].Computer applications(in China), 2014,34 (09): 2577-2580.

[44] Zhang Dawei.[J].Police Technology(in China), 2017 (06): 38-41.

[45] Zhao Rui.Multi-label parallel evolution supernetwork under the Spark platform [D]. Chongqing University of Posts and Telecommunications(in China), 2017.

[46] Zheng Weiping, Yang Xudong. On the Public Security Data Label System Construction [J]. Guangdong Public Security Technology(in China), 2019,27 (02): 9-12.

[47] Zhou Rui, Huang Jing, Fan Jiaqi. Research on the Construction of Big Data Portrait of Public Health Incident [J]. E-government affairs(in China), 2020(06):12-20.DOI:10.16582/j.cnki.dzzw.2020.06.002.

[48] Zhu Zhenhua, Yu Xiaoyun, Li Chao. Application Analysis and Research of Personnel Background Label Based on Public Security Big Data [J]. Computer Knowledge and Technology (in China), 2018,14(21):28-30.DOI:10.14004/j.cnki.ckt.2018.2286.