Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8752022, 9 pages
https://doi.org/10.1155/2021/8752022

Research Article

K-Nearest Robust Active Learning on Big Data and Application in Epitope Prediction

Tianchi Lu

School of Mathematics and Statistics, Lanzhou University, Lanzhou, China

Correspondence should be addressed to Tianchi Lu; lutch17@lzu.edu.cn

Received 11 August 2021; Revised 10 September 2021; Accepted 17 September 2021; Published 11 November 2021

Academic Editor: Rajesh Kaluri

B-cells that induce antigen-specific immune responses in vivo produce large numbers of antigen-specific antibodies by recognizing subregions (epitopes) of antigenic proteins, in which they can inhibit the function of antigen protein. Epitope region prediction facilitates the design and development of vaccines that induce the production of antigen-specific antibodies. There are many diseases which are difficult to treat without vaccines. And the COVID-19 has destroyed many people's lives. Therefore, making vaccines to COVID-19 is very important. Making vaccines needs a large number of experiments to get labeled targets. However, obtaining tremendous labeled data from experiments is a challenge for humans. Big data analysis has proposed some solutions to deal with this challenge. Big data technology has developed very fast and has been applied in many areas. In the bioinformatics area, big data analysis solves a large number of problems, particularly in the area of active learning. Active learning is a method of building more predictive models with less labeled data. Active learning establishes models with less data by asking the oracle (human) for the most valuable samples to train models. Hence, active learning's application in making vaccines is meaningful that the scientists do not need to do tremendous experiments. This paper proposed a more robust active learning method based on uncertainty sampling and K-nearest density and applies it to the vaccine manufacture. This paper evaluates the new algorithm with accuracy and robustness. In order to evaluate the robustness of active learners, a new robustness index is designed in this paper. And this paper compares the new algorithm with a pool-based active learning algorithm, density-weighted active learning algorithm, and traditional machine learning algorithm. Finally, the new algorithm is applied to epitope prediction of B-cell data, which is significant to making vaccines.

1. Introduction

Big data analysis is a thriving field. The branch of big data analysis, artificial intelligence, has greatly promoted the team's understanding of life science in the field of bioinformatics [1, 2]. We can use machine learning to predict major disease problems for the benefit of human beings, such as vaccine manufacturing.

Now, people are fully aware of the importance of health. At the same time, with the development of the Internet and big data, many mobile applications with collaborative systems have been developed to detect people's health [3–5]. As we can see from previous work, many collaborative systems have begun to work with machine learning methods [3–5]. These systems can use machine learning to detect body states. At the same time, COVID-19 disease has destroyed many people's lives, so it is necessary to use such a system to detect whether people have COVID-19. And making vaccines to COVID-19 is also an emergency. Therefore, we choose the B-cell [6, 7] data which is so relative to the immune system. Antibodies inhibit the function of antigen proteins by identifying antigen epitope that can be seen as "vaccines," because B-cells are immune cells that can recognize antigens when producing antibodies. Therefore, predicting epitopes using B-cell data [6, 7] is important for the preparation of experimental vaccines. Since it is not difficult to get the specimen of B-cell, using the epitope prediction to the health collaborative systems is a good way to assess whether people are suffering from COVID-19. Several physical and computational methods [8–12] have been proposed to predict epitopes. In the physical methods, the features used are limited to those associated with the target

amino acid sequence, so the representations of these models are inadequate [6]. And using the physical methods to predict epitope requires tremendous experiments which need labor, establishment, and money. Although the computational methods have achieved better, it still requires tremendous samples to train. Therefore, these methods are expensive.

There are several ways to cope with the big data problem when reducing the burden of data and experiments. Dimensionality reduction [1] is one of the most important methods to reduce the complexity of models and select the most important variables. However, dimensionality reduction still requires tremendous samples. Active learning [13] is a solution to this problem. Active learning which aims to reduce the number of samples required by asking the oracle is a subfield of machine learning with the same name [14] in educational literature. And the area of active learning is booming: many active learning methods [15, 16] have been proposed. These algorithms are based on uncertainty sampling. In addition to uncertainty sampling, many sampling processes for active learning are proposed. Based on label changes [17], committee queries [18], representative changes [19], and density-based sampling [20, 21] are some of the processes. More importantly, the active learning method has been successfully applied to speech recognition [22], information extraction [23], and bioinformatics [24–26].

We believe that using active learning is of great significance to predict epitope, and this paper mainly concentrates on the uncertainty sampling [27, 28] and density sampling method [20, 21]. The uncertainty sampling method usually selects the outliers as the most uncertain and informative samples to ask the oracle. Outliers are not so valuable and may result in less robust classifiers when new samples are added to the training data. To solve this problem, density-weighted sampling has been proposed. Density-weighted sampling [20, 21] is a good way to solve the outlier problem. But the density-weighted method does not provide the same information as the uncertainty sampling method. Recently, some methods have developed new loss functions by integrating uncertainty sampling and K-nearest density weighting to improve the performance of active learning [29–31]. However, these methods may still cause loss of information, just like the density-weighted method. And calculating the density of samples in the pool samples is difficult since the computing complexity is great when there are many samples [32]. In order to use the most valuable data and make more robust query strategies without high complexity, this paper establishes a new algorithm. Specifically, the work uses uncertainty sampling to find the most informative points firstly, then uses K-nearest density in the uncertainty data with L_1 norm (Manhattan distance) to eliminate outliers to improve pool-based active learners' robustness. This paper calls the new algorithm K-nearest robust active learning (KRAL). Compared to the density-weighted method like SUD [31], the KRAL is with less complexity. This is because not many most uncertainty samplers are generated in each step; calculating density in this dataset does not result in computing complexity being too high. At the same time, using the K-nearest density method, we eliminate the outliers, which guarantees the maximization of information utilization and does not increase computing complexity excessively.

Our proposal is to make a new algorithm which predicts the epitope with less labeled data and higher accuracy when compared to the existed pool-based active learning and density-weighted active learning algorithms in epitope prediction problem. Hence, this paper uses B-cell data with epitope to do the experiments. The data comes from the immune epitope database (IEDB) which is a public database of immune epitope [7]. By experimenting and comparing the KRAL with pool-based active learning and density-weighted method on B-cell data, we finally get a more accurate and robust model with less complexity. Therefore, the results of this study may be helpful in the production of the COVID-19 vaccine.

2. Data and Methodology

2.1. Data and Task Description. The world is suffering from a pandemic in which COVID-19 has destroyed a large number of people's lives. Substances that mimic the structure and function of epitopes can be thought of as "vaccines" of organisms designed to induce specific antibodies in vivo. Therefore, the B-cell data [6] is selected for this study. B-cells are immune cells that recognize antigens when producing antibodies. Antibodies can inhibit the function of antigen proteins by binding to antigen epitope regions. Hence, it is very helpful to find a good prediction model of epitope for this problem. There are some physical methods to predict the epitope. For instance, the three-dimensional structural analysis of antibody-antigen complexes by X-ray [9] or nuclear magnetic resonance (NMR) spectroscopy [10] is considered to identify the epitope.

But these methods are quite expensive and require a lot of time and labor to predict epitope. Recently, various big data analysis methods were proposed based on machine learning [11, 12]. Under this circumstance, the performance of epitope prediction has improved by machine learning methods. But we still need a lot of data for training, which is still expensive. Hence, the task is still challenging for humans. Next, we describe this task in detail.

The data and variables description:

Independent variables:

(i) start_position: start position of peptide

(ii) end_position: end position of peptide

(iii) chou_fasman: peptide feature, β turn

(iv) emini: peptide feature, relative surface accessibility

(v) kolaskar_tongaonkar: peptide feature, antigenicity

(vi) parker: peptide feature

(vii) isoelectric_point: protein feature

(viii) aromaticity: protein feature

(ix) hydrophobicity: protein feature

(x) stability: protein feature

Dependent variable:

(i) Antibody valence (target value)

The task is a binary classification problem with 10 independent variables, and the target was antibody valence, where 0 stands for negative and 1 stands for positive. There are 14387 samples in the data. The structure of the dataset is shown in Figure 1 and Table 1. Figure 1 and Table 1 illustrate that about 3/4 samples are negative and 1/4 are positive. From the skewness and kurtosis from Table 1, we can see that some of the independent variables do not follow the normal distribution, and some are sparsely distributed. Particularly, the end_position, start_position, and emini are not obeying the normal distribution. And some others, like the hydrophobicity, are sparsely distributed. Therefore, the dataset may have some outliers that may affect the performance of active learning algorithms. Therefore, traditional machine learning and active learning methods may not work well.

2.2. Methodology Description. In this paper, we propose a new big data analysis method to predict epitope, and B-cell data were used to establish the model. The detailed steps of this work are shown in Figure 2. More specifically, this paper uses KRAL to predict targets and incorporates the new algorithm with traditional pool-based active (PBL) learners, density-weighted active learning method (SUD), and basic algorithms (random forest [33] and SVM [34]) with random selection (RS) in both accuracy and robustness. In order to evaluate the active learners' robustness, this paper designs a new index called sequential robust index (SRI).

3. Active Learning Process

This paper is interested in big data and pool-based active learning based on uncertainty sampling [15]. That is, active learners have the least confidence in the samples with the greatest uncertainty, while pool-based active learners have two-stage samples. There are a small number of labeled samples and a large number of unlabeled samples. Pool-based active learners require oracle to provide the most uncertainty samples and add them to the labeled samples for the next training. The full algorithm is illustrated as follows [16]. The algorithm results are shown in Algorithm 1.

Pool-based active learning:

4. Uncertainty Measures

There is too much useless information when dealing with big data. Therefore, choosing the sample with the most useful information is important. In the uncertainty sampling scheme, the unlabeled sample with the largest uncertainty is considered the one with the largest amount of information. Therefore, it is significant to find a good evaluation method of measurement sample uncertainty.

The well-known entropy [27] has been widely used in previous studies in evaluating uncertainty [35, 36].

$$\mathbf{H}(\mathbf{x}) = -\sum_{\mathbf{y}\in\mathbf{Y}} \mathbf{P}(\mathbf{y} \mid \mathbf{x})\mathbf{logP}(\mathbf{y} \mid \mathbf{x}), \tag{1}$$

where $P(y \mid x)$ is the a posteriori probability, target (label) $\mathbf{y} \in \mathbf{Y} = \{y_1, y_2, \cdots y_k\}$. $\mathbf{H}(\mathbf{x})$ is the uncertainty measurement function based on the entropy estimation of the classifier's posterior distribution.

Entropy is a baseline method for measuring uncertainty, which involves a large amount of information. Therefore, this paper uses the entropy as the uncertainty sampling measurement.

5. K-Nearest Robust Active Learning

Although entropy contains the most useful information, it has some drawbacks. Entropy is to find the nearest point to the classification boundary, that is, outliers are usually used as the sample with the least confidence. Outliers may contain too much noise, resulting in poor robustness of the model. Therefore, dealing with outliers is a feasible way to improve the model's performance.

5.1. K-Nearest Neighbor Classification. K-Nearest Neighbor Classification (KNN) [32] learners are the basic way to deal with classification problems. KNN is characterized by estimating sample density using a distance function. Therefore, using K-nearest density can help us find outliers and pass them out of the training data. However, the KNN model is a method with great complexity. Therefore, how to use the K-nearest neighbor algorithm in active learning is a challenge.

5.2. Distance Function. There are many distance functions. Hence, selecting a fit distance function is fundamental to estimate the K-nearest density of samples. Considering the effect of outliers and the complexity of big data, this paper uses the Manhattan distance to calculate the K-nearest density.

Manhattan distance:

$$d_{12} = \sum_{k=1}^{n} |X_{1k} - X_{2k}|. \tag{2}$$

The Manhattan distance method is not affected so much by outliers. It evaluates whether two points are close or not. And the Manhattan distance is also not so complex. Therefore, calculating the Manhattan distance will not add too much complexity in big data. Hence, using Manhattan distance to calculate the distance between vectors is a good choice.

5.3. K-Nearest Robust Active Learning. Using the K-nearest classification to calculate the density of samples can help us to find the outliers. The density function $\mathbf{Den}(\mathbf{X_i})$ is

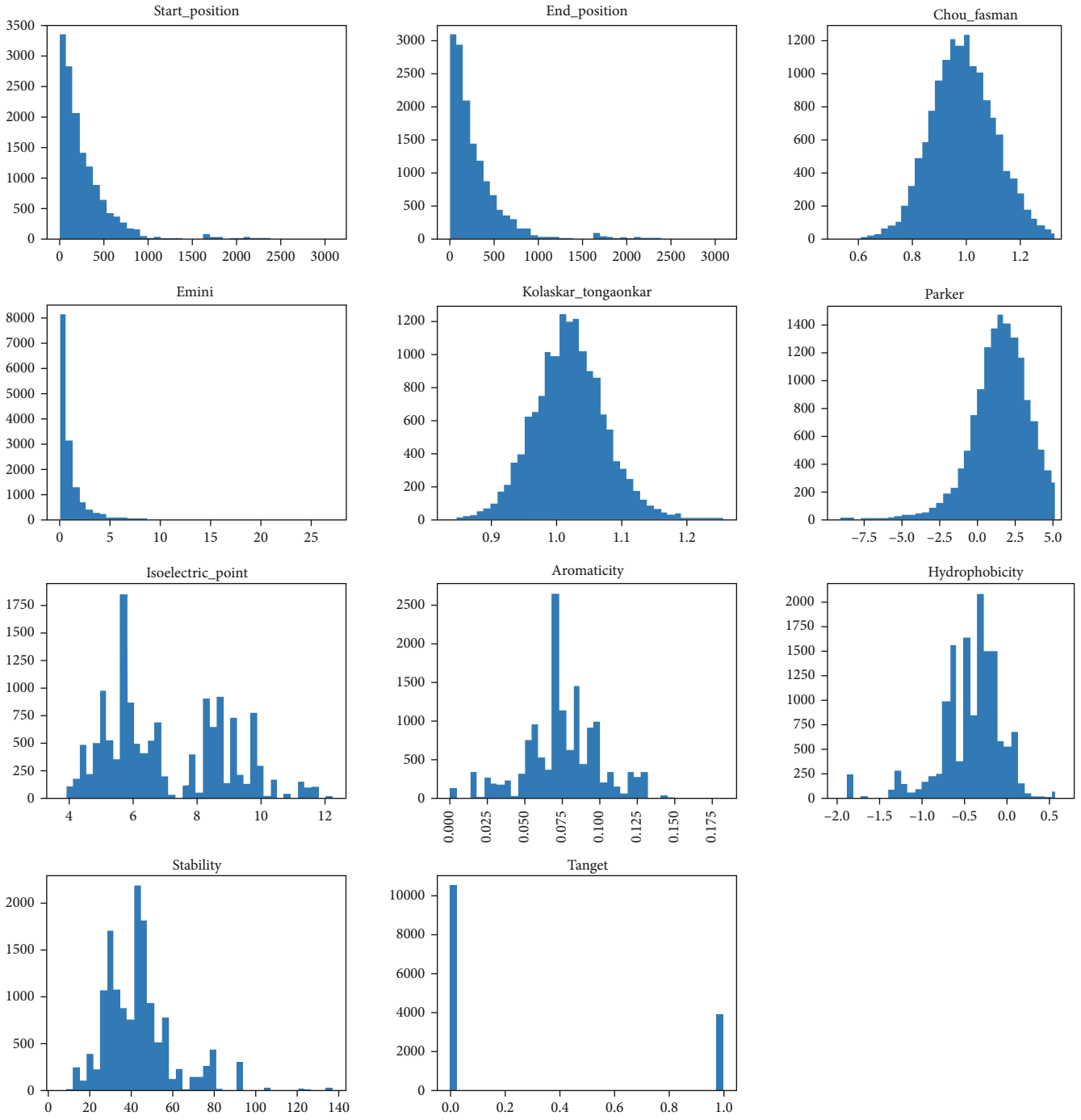

Figure 1: The data describing.

defined as

$$\mathbf{Den}(\mathbf{X_i}) = \frac{\mathbf{1}}{\sum_{\mathbf{J=1}}^{\mathbf{m}} \mathbf{d}(\mathbf{X_i}, \mathbf{X_j})}, \tag{3}$$

where $\mathbf{m}$ is the number of uncertainty samples. $\{\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_m}\}$ are the most $\mathbf{m}$th uncertainty samples. In the KRAL, the $\mathbf{K}$ is defined as the $\mathbf{m}$, which means that we calculate the distance among the most uncertainty samples at every step. From the form of density function, the sample with the smallest density function is the point which needs to be threaded out. Compared to a naive opinion which is to apply the K-nearest procedure like SUD for all unlabeled data, the new density functions do not add much complexity to the algorithm. This is because calculating the K-nearest density on a big data will increase the complexity greatly. And it is difficult to decide the $\mathbf{K}$ value because the whole unlabeled data is concluded too much samples. Hence, we cannot simply use $\mathbf{m}$ (number of samples) as the $\mathbf{K}$.

Table 1: An error and statistical analysis to the data.

	Descriptive statistics							
	N	Mean	Std. deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. error	Statistic	Std. error
Start_position	14387	297.68	353.741	125133.014	3.009	.020	11.607	.041
End_position	14387	308.09	353.733	125127.245	3.005	.020	11.574	.041
Chou_fasman	14387	.994705915000000	.124772254000000	.016	.248	.020	.398	.041
Emini	14387	1.059787725000000	1.621931429000000	2.631	5.051	.020	40.411	.041
Kolaskar_tongaonkar	14387	1.021188364000000	.053804291800000	.003	.186	.020	.380	.041
Parker	14387	1.767136582000000	1.968984865000000	3.877	-.362	.020	1.266	.041
Isoelectric_point	14387	7.067471661000001	1.888708170000000	3.567	.439	.020	-.915	.041
Aromaticity	14387	.075726787200000	.025767473200000	.001	-.131	.020	.570	.041
Hydrophobicity	14387	-.406096679000000	.394618135000000	.156	-.706	.020	3.058	.041
Stability	14387	43.703902170000000	16.682362480000002	278.301	1.366	.020	3.248	.041
Valid *N* (listwise)	14387							

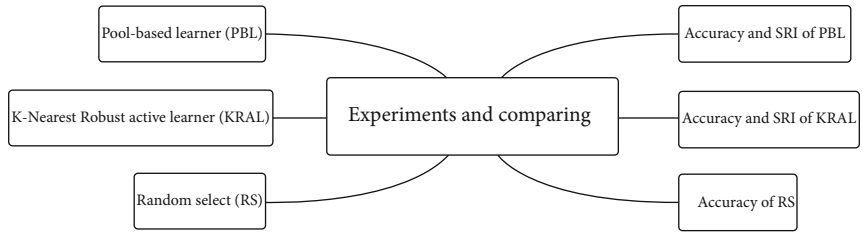

Figure 2: The step of this paper's work.

Require: A set of labeled samples L, a set of unlabeled samples U
  while Termination condition not satisfied do
    Train a classifier $\varphi c(\cdot|L)$ based on labeled samples;
    for i =1 : |U| **do**
      Calculate the uncertainty(Entropy) of the sample, $Un(x_i^u)$;
      Select the top nth uncertainty sample y_i as a new set N to query;
      L=L∪N;
      U=U-N;
    end for
  end while

Algorithm 1: Pool-based active learning process

Therefore, how to choose the **K** is a challenge. And the most uncertainty samples are more like to be outliers. Hence, calculating the density among the most uncertainty samples is a good idea. Hence, this paper can exclude outliers from the least confidence level sample with K-nearest density.

Then, we detail the new algorithm steps. The new algorithm (KRAL) uses entropy sampling in the first step to select the most indeterminate sample, which is the same as pool-based active learning. Next, KRAL uses Manhattan distance to calculate the K-nearest density in the most uncertain samples at the next step. Then, the KRAL selects the sample with the lowest density for threading, because the big data is so complex and large that many methods cannot be used to cope with big data. From the form of KRAL, we can see that KRAL both consider the computing complexity and accuracy. Hence, the KRAL can be applied in big data analysis. And the new full algorithm of KRAL is illustrated as follows. The algorithm results are shown in Algorithm 2.

K-nearest robust active learning:

6. Experiments and Numeric Research

In this section, experimental and numerical studies are performed using B-cell data. More specifically, this paper compares the accuracy and robustness of KRAL and pool-based active learning.

6.1. Sequential Robust Index (SRI). There is no method that has been proposed to evaluate the robustness of active learning when adding new samples to training data. In order to

Require: A set of labeled samples L, a set of unlabeled samples U
  while Termination condition not satisfied do
    Train a classifier $\varphi c(\cdot|\mathbf{L})$ based on labeled samples;
  for i =1 : |U| **do**
    Calculate the uncertainty(Entropy) of the sample, $\mathbf{Un}(\mathbf{x_i^u})$;
    Select the top nth uncertainty sample $\mathbf{y_i}$ as a new set N ;
    Calculate the density function $\mathbf{Den}(\mathbf{y_i})$ in set N;
    p= **argmin** $\mathbf{Den}(\mathbf{y_i})$;
    N=N-p;
    use N's samples to query the oracle;
    L=L∪N;
    U=U-N;
  end for
  end while

Algorithm 2: K-nearest robust active learning process

evaluate the robustness of the algorithms, a new robustness index (SRI) for the sequence of robustness evaluation indexes is presented. The sequential robust index is defined as

$$\sum \boldsymbol{I}(\boldsymbol{a}_i - \boldsymbol{a}_{i-1} < \mathbf{0}) \sum |\boldsymbol{a}_i - \boldsymbol{a}_{i-1}| \boldsymbol{I}(\boldsymbol{a}_i - \boldsymbol{a}_{i-1} < \mathbf{0}), \quad (4)$$

where the $\mathbf{a_i}$ is the accuracy of one-step test data and $\mathbf{I}(\mathbf{x})$ is the indication function. When $x < 0$, $\mathbf{I}(\mathbf{x}) = 1$; otherwise, $\mathbf{I}(\mathbf{x}) = 0$.

We expect that when new samples are added to the training set, the prediction accuracy of the test set will increase. Through this way, we can reduce the computational complexity when facing big data. However, sometimes, adding new samples into the training set in active learning process will result in a lower prediction accuracy. The SRI measures the number of times predictive accuracy decreases and the total amount of decline when new samples are added to the training set. Because the fewer times the accuracy is reduced and the fewer the accuracy is reduced when adding samples into training data, the more valuable the data is added to the training set each time. Hence, it can be seen from the form of SRI that the smaller the index, the better the model. We can see that if a good query strategy is stable, the new data it queries will make the proactive learning prediction accuracy increasing. However, if a query strategy is unstable, the queried data may reduce the prediction accuracy, so SRI can measure the stability of a query strategy greatly at some step, that is, SRI evaluates the robustness of the query strategy. Therefore, the SRI can estimate the robustness of active learning during the query process. And the computation complexity of SRI is not high. So SRI can very deal with big data.

6.2. Experimental Settings. We use random forest (RF) and support vector machine (SVM) as base learners. And the query strategy is based on maximum entropy. Cross-validation is a good way to examine the performance of models in big data analysis. Therefore, in order to ensure the rationality of the experiment, we randomly select samples as labeled data by cross-validation and repeat the experiment 100 times and use the mean value to record in results. To be more specific, we randomly divide the data into 50 parts using 50-fold cross-validation and randomly select one of them as training set and the rest as pools for active learning queries. The data is a public dataset (IEDB) [7], which we will use for epitope prediction. And we use the pool samples as test set at every query step.

We compare KRAL with pool-based active learning and SUD. And our evaluation metrics are the test set accuracy, SRI, and the running time. Among them, test set accuracy is used to directly measure the effectiveness of several methods, SRI is used to evaluate the query robustness of several methods, that is, to evaluate the stability of the query, and running time is to evaluate the computational complexity of the model. We mainly compare the effectiveness and computational complexity of every algorithm.

In every query step, we let the most uncertain dataset includes 40 samples. Under this circumstance, we continue our experiments. And my computer setting is GPU: RTX 3060 and CPU: 16G, I7, 11th generation.

The IDE is Spyder.

7. Result and Analysis

This paper records the accuracy and SRI when the number of samples increases. The results are recorded in Figures 3–6. Figures 3 and 5 record the accuracy of each learner, and Figures 4 and 6 record the SRI of each active learner. From Figures 3–6, we can see the results of each model: random selection sampling is the weakest in both random forests and support vector machine models. As the number of samples increases, the sensitivity of random selection sampling decreases. In the SVM model, adding new samples to the training data does not significantly improve the accuracy. Both pool-based active learning, KRAL and SUD methods, improve the performance of basic learners. Figures 3–6 show that when new samples are added, the active learning's accuracy is higher than the basic learners. Therefore, using the active learning method can reduce the complexity when coping with big data. And when the basic learner is random forest, the performance of KRAL is 12.1% better than that of the basic learner. Therefore, the effectiveness of KRAL was

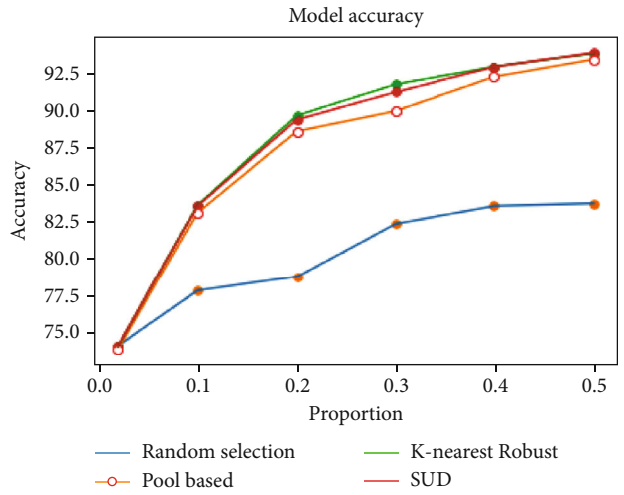

Figure 3: The accuracy in the random forest model.

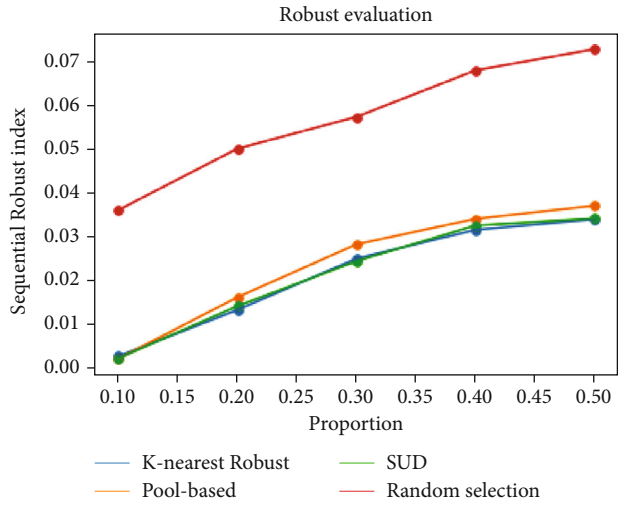

Figure 4: The robustness in the random forest model.

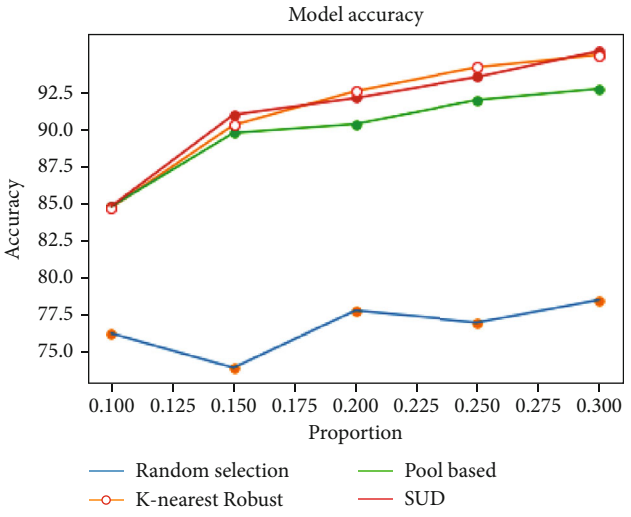

Figure 5: The accuracy in the SVM model.

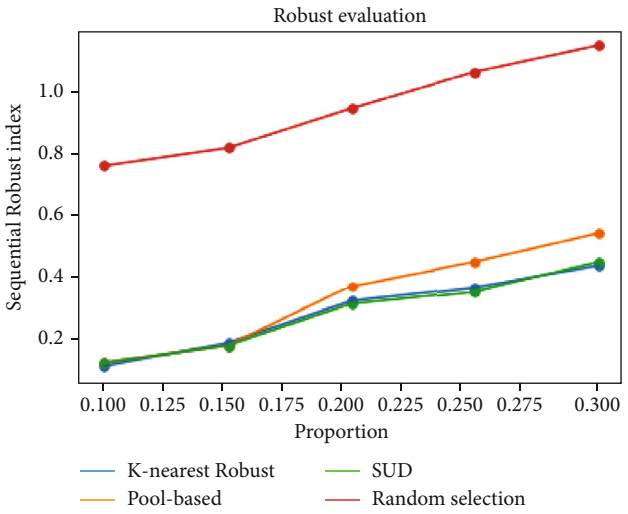

Figure 6: The robustness in the SVM model.

examined. So active learning methods can be used to reduce the computational complexity in big data analysis and improve the accuracy.

Meanwhile, the SRI of KRAL and SUD is smaller than pool-based active learning and random selection strategy, indicating that KRAL and SUD are more robust than pool-based active learning. And the SRI among the active learners is much smaller than random selection. More specifically, the KRAL algorithm achieves at least 6% and 15% higher in SRI evaluation and the SUD algorithm achieves at least 5.8% and 15.1% higher in SRI evaluation than pool-based active learning algorithm in random forest and SVM. And KRAL and SUD algorithms are more accurate than pool-based active learning algorithm in SVM and RF models. When the sample scale is 5/10, the learning accuracy of KRAL is 0.5% higher than that of pool-based active learners in both two basic learners (RF and SVM). Therefore, SUD and KRAL can use less data to establish a better model than pool-based active learners.

And we can see that the prediction accuracy of RF and SVM is different. This is because RF is an ensemble learning method, and its base learner is a decision tree. A decision tree is not a linear regression or classification method, it can be applied to different types of datasets. At the same time, the use of ensemble learning and certain randomness make RF have stronger generalization ability. In this experiment, we use linear SVM, which form is simple and cannot deal with the complex data structure. And SVM does not use ensemble methods. Therefore, the effect of SVM is weaker than that of RF in this experiment.

However, if we only use the accuracy and SRI to evaluate SUD and KRAL, we cannot tell the difference between the two algorithms. However, as we mentioned, the SUD uses the whole unlabeled data to calculate its K-nearest density. But using whole unlabeled data to calculating density will cause the increase of computational complexity. In order to evaluate the complexity, we use the time consuming of every algorithm. Figure 7 shows that the KRAL's complexity is strongly lower than SUD. SUD is the time consumed

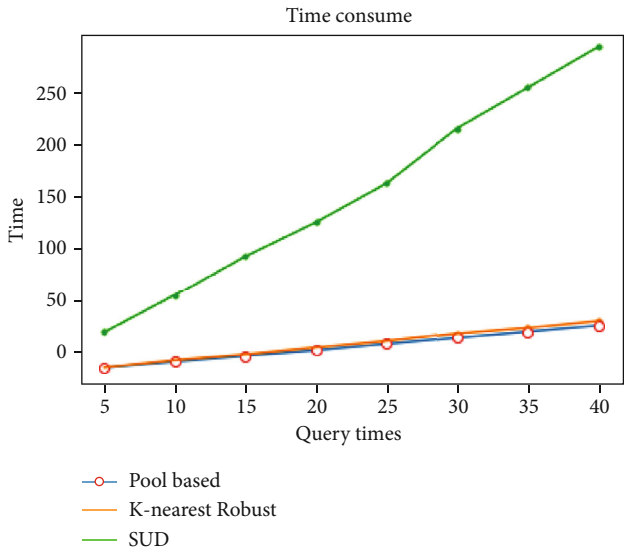

Figure 7: Time consumption in random forest.

when compared to the KRAL and the basic pool-based active learners, which means KRAL is more fit to deal with the big data problem than the SUD.

It can be seen from the experiment that the performance of the KRAL algorithm is better than pool-based active learning in accuracy and robustness and SUD in computational complexity. Therefore, a more robust and accurate method with less computational complexity is obtained, especially when KRAL is applied to outlier-sensitive models like SVM. This is because KRAL can select information data instances with fewer outliers. And the RF is not so sensitive to the outliers, which may reduce the effect of KRAL and SUD. Therefore, considering accuracy, robustness, and computation complexity in big data analysis, this paper uses the KRAL algorithm with random forest to predict B-cell data epitopes. The accuracy of the model obtained in this paper is 93.8%, and only 4/10 of the samples are used.

8. Conclusion and Discussion

The contribution of this paper to big data analysis is to propose a new more robust active learning method with higher accuracy and a new active learning robustness evaluation metric SRI. The new algorithm can also reduce the complexity of density-weighted pool-based active learners like SUD when facing the big data. And the effectiveness and robustness of KRAL, SUD, and pool-based active learning are evaluated experimentally by the SRI. Through the experiments, a more robust and accurate algorithm with less complexity is obtained. Apart from the computational complexity, KRAL has some advantages in big data area when compared with the SUD algorithm. More specifically, KRAL eliminates outliers by estimating sample density for better performance. However, SUD only uses a new loss function to change the structure of the model. Hence, KRAL has greater potential. This is because scholars can change the proportion of deleted samples before adding them to the training data. Specifically, using a dynamic greedy algorithm with a reasonable loss function to improve KRAL's performance is a prospective direction. Therefore, when the basic learner is not sensitive to outliers, the algorithm can achieve better results. However, SUD cannot use this method to improve performance. But the KRAL also has some disadvantages: KRAL still uses the uncertainty query strategies for searching the most valuable samples. However, this may be not fit in many areas such as the natural language processing (NLP). Therefore, changing the query strategies to fit these areas is a good direction. In the future, the author will look for a good loss function to improve the performance of KRAL and look for some new query strategies for active learning and make more contributions in big data and artificial intelligence area. To be more specific, the author will devote himself into the NLP area and find more suitable query strategies to let the active learning method more effective in such as Neural Machine Translation (NMT) problem. And the author will conduct some research in bioinformatics to find some cure to kinds of diseases.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author declares that there are no conflicts of interest concerning the publication of this paper.

Acknowledgments

The author would like to thank Jiatong Shi, who gives some specific and significant suggestions to this paper.

References

[1] D. Guru and S. Perumal, "Approaches towards blockchain innovation: a survey and future directions," *Electronics*, vol. 10, no. 10, pp. 1219–1219, 2021.

[2] N. Deepa, Q. V. Pham, and D. C. Nguyen, "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020.

[3] A. R. Javed, M. U. Sarwar, M. O. Beg, M. Asim, T. Baker, and H. Tawfik, "A collaborative healthcare framework for shared healthcare plan with ambient intelligence," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–21, 2020.

[4] M. U. Sarwar and A. R. Javed, "Collaborative health care plan through crowdsource data using ambient application," in *2019 22nd International Multitopic Conference (INMIC)*, pp. 1–6, Islamabad, Pakistan, 2019.

[5] A. Arj, B. Lgf, and B. Aaf, "Automated cognitive health assessment in smart homes using machine learning," *Sustainable Cities and Society*, vol. 65, 2021.

[6] T. Noumi, S. Inoue, H. Fujita et al., "Epitope prediction of antigen protein using attention-based LSTM network," *Journal of Information Processing*, vol. 29, pp. 321–327, 2021.

[7] R. Vita, J. A. Overton, J. A. Greenbaum et al., "The immune epitope database (IEDB) 3.0," *Nucleic Acids Research*, vol. 43, no. D1, pp. D405–D412, 2015.

[8] H. M. Regenmortel, "The concept and operational definition of protein epitopes," *Philosophical Transactions of the Royal Society of London*, vol. 323, no. 1217, pp. 451–466, 1989.

[9] J. Rux, "Type-specific epitope locations revealed by X-ray crystallographic study of adenovirus type 5 hexon," *Molecular Therapy the Journal of the American Society of Gene Therapy*, vol. 1, no. 1, pp. 18–30, 2000.

[10] M. Mayer and B. Meyer, "Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor," *journal of the american chemical society*, vol. 123, no. 25, pp. 6108–6117, 2001.

[11] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, "BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes," *Nucleic Acids Research*, vol. 45, no. W1, pp. W24–W29, 2017.

[12] H. Singh, H. R. Ansari, and G. P. S. Raghava, "Improved method for linear B-cell epitope prediction using antigen's primary sequence," *PLoS One*, vol. 8, no. 5, 2013.

[13] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, no. 4, pp. 319–342, 1988.

[14] C. C. Bonwell, *Active Learning: Creating Excitement in the Classroom. ERIC Digest*, ERIC Clearinghouse on Higher Education, Washington, DC, 1991.

[15] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," Heidelberg: Springer Verlag, Berlin, 1994.

[16] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[17] P. Juszczak and R. Duin, "Selective sampling based on the variation in label assignments," in *ICPR 2004. Proceedings of the 17th International Conference on. 2004*, 2004.

[18] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," Association for Computing Machinery, New York, NY, USA, 1992.

[19] F. Sebastiani, "Representative sampling for text classification using support vector machines," *Lecture Notes in Computer Science Advances in Information Retrieval*, vol. 2633, pp. 393–407, 2003.

[20] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2001.

[21] S. C. H. Hoi, R. Jin, and J. Zhu, "Batch mode active learning and its application to medical image classification," Association for Computing Machinery, New York, NY, USA, 2006.

[22] X. Zhu, *Semi-Supervised Learning with Graphs*, PhD thesis, Carnegie Mellon University, 2005.

[23] B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," *Proceedings of the NIPS Workshop on Cost-Sensitive Learning.*, vol. 1, 2008.

[24] A. W. Naik, J. D. Kangas, D. P. Sullivan, and R. F. Murphy, "Active machine learning driven experimentation to determine compound effects on protein patterns," *eLife*, vol. 5, 2016.

[25] T. Nakano, S. Takeda, and J. Brown, "Active learning effectively identifies a minimal set of maximally informative and asymptotically performant cytotoxic structure–activity patterns in nci-60 cell lines," *RSC Medicinal Chemistry*, vol. 11, no. 9, pp. 1075–1087, 2020.

[26] M. Hafner, M. Niepel, K. Subramanian, and P. K. Sorger, "Designing drugresponse experiments and quantifying their results," *Current Protocols in Chemical Biology*, vol. 9, no. 2, pp. 96–116, 2017.

[27] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] C. Campbell, N. Cristianini, and A. J. Smola, "Query learning with large margin classifiers," Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2000.

[29] Y. J. Gu, D. Zydek, and Z. Jin, "Active learning based on random forest and its application to terrain classification," in *Progress in Systems Engineering*, Advances in Intelligent Systems and Computing, volume 366, Springer, Cham, 2015.

[30] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1323–1331, 2010.

[31] J. Zhu, H. Wang, and T. Yao, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," Coling 2008 Organizing Committee, Manchester, UK, 2008.

[32] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1953.

[33] Breiman, "Random forests," *MACH LEARN*, vol. 45, no. 1, pp. 5–32, 2001.

[34] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[35] M. Tang, X. Luo, and S. Roukos, "Active learning for statistical natural language parsing," *USA: Association for Computational Linguistics.*, pp. 120–127, 2002.

[36] J. Zhu and E. H. Hovy, "Active learning for word sense disambiguation with methods for addressing the class imbalance problem," in *Conference on Emnlp-conll*, DBLP, 2007.