Active Ensemble Learning: Application to Data Mining and Bioinformatics

Hiroshi Mamitsuka* and Naoki Abe

Internet Systems Research Laboratories, NEC Corporation, Kawasaki, 216-8555 Japan

SUMMARY
This paper describes a new set of learning procedures which have been proposed by the authors. The method combines active learning and the accuracy enhancement techniques of bagging and boosting, and may be called active ensemble learning. Any of these procedures achieves highly accurate learning by iteratively selecting (querying) a small amount of data with large information content from a data space or database. This paper describes not only the technical aspect of the method, but also the results of application to two real problems, namely, active planning of biochemical or molecular biological experiments in immunology, and customer segmentation from a large-scale body of data in the CRM (customer relationship management) field. It is demonstrated that the proposed methods can achieve greater data efficiency and prediction accuracy than conventional methods. © 2007 Wiley Periodicals, Inc. Syst Comp Jpn, 38(11): 100–108, 2007; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.10355

Key words: active learning; bagging; boosting; data mining; experimental planning.

1. Introduction
By ensemble learning is generally meant the procedure in which multiple hypotheses are prepared on the basis of given data, and the appropriate combination of these hypotheses, followed by prediction, is used to improve learning performance compared to learning by a single hypothesis.

The idea of improving prediction accuracy by combining multiple hypotheses has been proposed since the 1980s [7, 9]. Two methods have been attracting interest since the mid-1990s, namely, bagging [3] and boosting [8], which differ radically from the methods proposed earlier in that the data are selected strategically in forming hypotheses.

Bagging is a very simple algorithm in which the data are resampled uniformly and at random, and prediction is performed by majority decision of the output hypotheses from multiple learning. The effectiveness of the method in improving prediction accuracy has been demonstrated by using real data. Boosting also resamples the data, as in bagging, but also has a further feature: the distribution is not uniform, and the probability distribution is increased for the data with low prediction accuracy in the previous learning. In addition, prediction is made by weighted majority decision of multiple hypotheses. The prediction accuracy is guaranteed theoretically. There have been many theoretical studies of these methods, as well as many recent applications to real problems. In the case of bagging, its effectiveness in improving prediction accuracy has been demonstrated.

By active learning is meant a learning configuration in which the learner himself selects the data. As a result of this involvement, the data efficiency can be improved and the amount of computation can be reduced. Active learning is broadly divided into methods in which data with higher
information content are selected by an information-theoretic approach, and the efficient query learning by an algorithm-theoretic approach [2].

A typical method of the former class is called “query by committee” (below called mass query learning) [19]. In this method, the same data are input into multiple learning algorithms, and the data for which the predictions by multiple hypotheses are most scattered are defined as the data with high information content and are used as the new query points. The effectiveness of the method has been shown mostly from the theoretical standpoint. A problem with this method is that the same data must be input. Thus, the learning algorithm to be used is restricted to random algorithms that can output different hypotheses for the same data.

This paper notes the similarity between ensemble learning and mass query learning, and describes a set of active learning procedures which can be derived by combining these two approaches. In particular, these methods are shown to be very useful by the experimental investigation of two real problems, namely, dynamic planning of biochemical and molecular-biological experiments in imitation of two real problems, namely, dynamic planning of.

active learning procedures which can be derived by combining these two approaches. In particular, these methods are shown to be very useful by the experimental investigation of two real problems, namely, dynamic planning of biochemical and molecular-biological experiments in immunology, and data mining from a large-scale body of data.

2. Query by Bagging/Boosting (Qbag/Qboost)

Mass query learning has the feature of ensemble learning that multiple hypotheses are learned in order to select the query points. The authors have noted this feature of ensemble learning and have proposed a new active learning procedure called query by bagging/boosting (Qbag/Qboost) [1]. In the proposed method, bagging/boosting, which are powerful accuracy enhancement techniques in ensemble learning, are combined with mass query learning, which can eliminate the constraints on the learning algorithm which existed in the original mass query learning.

In Qbag/Qboost, the multiple hypotheses treated by bagging or boosting are used in mass query learning instead of multiple hypotheses learned by using the same data. By using this process, the learning algorithms that can be used gain practical versatility and their accuracy is also enhanced.

Figure 1 shows the pseudo-code of Qbag. The Qbag algorithm is outlined as follows. Bagging is applied to the input data. The point at which the variance of the values predicted by the obtained multiple hypotheses is maximized is defined as a point with high information content and is used as the next query point. The resulting function value is then stored as accumulated data. After repeating this process, the final prediction value is obtained by majority decision or by taking the average of the final hypotheses.

This approach, in which data labels are not utilized and a point with low prediction reliability is defined as a point with high information content to be used as a new selection point, belongs to the category of uncertainty sampling [12], which is applied in the field of information retrieval. Furthermore, the approach of querying the sample

Algorithm: Query-by-Bagging

Input: number of trials: \( M \); learning algorithm at lower level: \( A \); number of resamplings in each trial: \( T \); number of candidate inquiry points: \( R \);

number of selected points: \( D \); set of labels: \( Y \);

candidate query points in \( i \)-th trial: \( C_i \);

stored selected points in \( i \)-th trial: \( S_i \)

Initialization:

Initial sample \( S_1 = \{(x_1, y_1), \ldots, (x_D, y_D)\} \) is selected at random.

For \( i = 1, \ldots, M \)

1. For \( j = 1, \ldots, T \)
   i. Resampling is applied from \( S_i \) with uniform distribution, and subsamples \( S'_j \) of the same size as \( S_i \) are determined.
   ii. Each of \( S'_j \) is used as the input to operate \( A \), and hypothesis \( h_j \) is obtained.

End For

2. \( R \) of \( C_i \) are selected at random.

3. For all of \( x \in C_i \), the following margin \( m(x) \) is calculated on the basis of hypotheses \( h_1, \ldots, h_R \), respectively.

\[
m(x) = \max_1(x) - \max_2(x)
\]

where

\[
\max_1(x) = \max_{t \leq T} |\{t \leq T : h_t(x) = y\}|
\]

\[
\max_2(x) = \max_{t \leq T} |\{t \leq T : h_t(x) = y\}|
\]

\[
y_{\max}(x) = \arg \max_{t \leq T} |\{t \leq T : h_t(x) = y\}|
\]

4. \( D \) points \( \{(x_1^*, y_1^*), \ldots, (x_D^*, y_D^*)\} \) with the smallest \( m(x) \) \( (x \in C_i) \) are selected from \( C_i \), and are added to the stored selected data (= learned data) as follows.

\[
S_{i+1} = \text{append}(S_i, \{(x_1^*, y_1^*), \ldots, (x_D^*, y_D^*)\})
\]

End For

Output: The final hypothesis is defined as follows.

\[
h_{\text{for}}(x) = \arg \max_{t \leq T} |\{t \leq T : h_t(x) = y\}|
\]

where \( h_t \) \( (t = 1, \ldots, T) \) is the hypothesis in the final \((M-\text{th})\) trial.

Fig. 1. Query by bagging (Qbag).
with the greatest information content can be considered as a technique of “selective sampling,” in which useful data are selected from the given data. Thus, Qbag/Qboost can be considered as a selective sampling technique.

For results of applying Qbag/Qboost to the benchmark data, see Ref. 1. In the next and following sections, performance measures of Qbag such as its data efficiency and accuracy are investigated by application to real data.†

3. Application Example 1: Use of Experimental Planning by Active Learning in MHC Binding Peptide Identification

As the first direct application example of active learning, we consider the MHC binding peptide identification problem, in which the cost of acquiring sample labels is very high. The application of Qbag to experimental planning in order to solve the problem is described below.

The recognition of not-self in the immunological response starts from the binding of not-self cell-origin peptides (short amino-acid sequences) to the molecule called MHC. The determination of which peptide binds strongly to each MHC would be a great contribution to medicine and pharmacology, making possible such techniques as the treatment of difficult diseases by means of the immunological system [20].

There exist 20 different amino acids, resulting in $20^{10}$ different combinations in peptides of length 10, for example. In addition, the cost of experimentally assessing the binding ability of one peptide to one MHC is several tens of thousands of yen and a period of 1 to 2 months.‡ Consequently, it is impossible to determine (measure) the MHC binding ability label for a large number of combinations. In fact, the presently available MHC binding peptide database [6] contains at most several hundred data. The identification of MHC binding peptides by utilizing these data has been attempted in several studies, but the accuracy is still insufficient [10, 13].

In this context, the authors attempted to generate highly accurate hypotheses from a small amount of data by applying “active-learning experiment planning” using Qbag for a particular MHC, and repeating a two-step procedure, consisting of the selection of a small number of peptides and actual measurement of the binding abilities of the selected peptides. The method was tested in collaboration with Associate Professor Udaka of Kyoto University Graduate School of Science.

†Specifically, the procedure of active learning experiment planning is to repeat the following two steps.

1. Application of a Qbag step. Multiple hypotheses are generated by resampling the accumulated data. The candidate query points are prepared at random, and the point with the greatest variance of the predicted value is selected as the query point for the experiment.†

2. The binding ability of the peptide at the query point is measured experimentally and results are added to the accumulated data.

Figure 2 shows schematically the above method of active learning experiment planning.

In the experiment, the hidden Markov supervised learning algorithm [13] is used for learning at the lower level. Learning starts from an initial data set of slightly fewer than 200 items. The above procedures are repeated slightly fewer than 10 times, resulting in a data set of slightly less than 400 items.

The results are evaluated by comparing the final hypothesis and the results obtained by the existing method. The method used for comparison is not a method involving automatic learning, but is a method called the “library method” [21] which has been favorably evaluated as a prediction procedure for MHC binding peptides. However, the library method cannot directly predict the binding ability.

Consequently, in the evaluation, the binding abilities of the peptides that are predicted to have the highest binding ability by the respective methods are compared in order to see which prediction procedure gives peptides with a

†Since Qboost is limited to binary label problems and bagging is more robust to the noise than boosting, Qbag is applied to the real data.

‡It is possible, however, to investigate multiple peptides in parallel in the same period.

†Specifically, peptides of length 9 are considered. Five million candidate query points are prepared at random, and the 20 to 40 points with the greatest variances are selected each time.
greater binding ability. Specifically, 1,000,000 peptides were randomly generated. Their binding abilities were evaluated by the two methods, and 41 peptides were selected at random from the top 10,000. The selected peptides were actually synthesized, and the binding abilities were measured and compared experimentally.

Table 1. Precision of Qbag and Library method

<table>
<thead>
<tr>
<th></th>
<th>Qbag</th>
<th>Library method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average binding ability</td>
<td>3.08</td>
<td>2.78</td>
</tr>
<tr>
<td>Fraction of peptides with binding ability greater than 3</td>
<td>0.59</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>(24/41)</td>
<td>(15/41)</td>
</tr>
</tbody>
</table>

Figure 3 shows the distribution of the measured binding abilities. Table 1 shows the average binding ability of the peptides selected by two methods. It is known that the threshold governing whether or not the peptide is bound to MHC is approximately 3, in terms of the binding ability. Table 1 shows the percentage of peptides with the binding ability exceeding 3.

As is evident from Fig. 3 and Table 1, the peptides selected by active learning experiment planning using Qbag provided higher binding abilities than those selected by the library method. The result indicates that prediction with higher accuracy can be realized from a small number of experiments by applying Qbag, which improves the data efficiency.

4. Application Example 2: Efficient Data Mining from Large-Scale Data

As an example of selective sampling based on active learning, data mining from large-scale data is next considered, and the application of Qbag to such a problem is discussed.

In present-day business based on CRM (customer relationship management), classification problems such as customer segmentation and customer targeting are becoming very important. It is required in particular to provide the ability to perform these tasks automatically by data mining. Furthermore, with increasing access by customers over the Internet, it is becoming relatively easy to store large-scale customer data as a business resource, and it is desirable to acquire knowledge from such data.

![Figure 4. Application of Qbag to mining large-scale database.](image-url)
By large-scale data we mean a scale of data that exceeds the capacity of the main memory of an ordinary computer. In such a situation, the whole body of data cannot be handled by ordinary passive learning methods. Consequently, a conceivable approach is to select (at random, for example) the data from the large-scale data to be contained in main memory, and to attempt learning. In such a procedure, there is a possibility that data-efficient, highly accurate data mining may be realized by repeated active selection of data with a high information content, as shown schematically in Fig. 4.

In order to verify the validity of such an approach, the membership data of an actual ISP (Internet service provider) were used. Using approximately 800,000 member data, the prediction of cancelers was attempted by churn analysis. Each data record contained two attribute sets, namely, demographic data such as the sex, age, and occupation of the member, and long/short-term traffic data, such as the amount of sent/received e-mail and the number of content accesses, as well as a binary label representing canceler/contractor.

The C4.5 decision tree [17] was used as the learning algorithm at the lower level in Qbag, and five-fold cross-validation was performed. Figure 5 compares the accuracy and precision of Qbag and C4.5 only in terms of the learning curves.†

We see from Fig. 5(a) that Qbag provides a higher accuracy than C4.5 alone. The data efficiency is also high in Qbag, reaching the same accuracy with fewer data. In practical applications, precision is more important than accuracy, since it represents the percentage of predicted cancelers who actually cancel. It is evident from Fig. 5(b) that the precision of C4.5 alone is at most 35% when several tens of thousands of data are used. On the other hand, Qbag reaches a precision of approximately 55% by using several thousand data, which is a much smaller amount. The accuracy is also improved by approximately 60%, which implies high data efficiency.

### 5. Query by Bagging: Single Loop (QbagS)

It has been stated that Qbag realizes high data efficiency, since only data with high information content are retained. When the data scale is further expanded in data mining from a large-scale data set, for example, the scale of the selected data must inevitably be expanded. It is to be expected in such a situation that the strategy of retaining only learned hypotheses, not the selected data, will be effective. Consequently, the authors devised an algorithm called QbagS (query by bagging: single loop) [14], which is a simplification of Qbag.

Specifically, in each stage of QbagS, only the hypotheses learned for the input data are retained, and the data are not stored. Based on multiple hypotheses obtained in the past, the point with the largest variance of the predicted values is defined as the next query point, that is, the input data. The final prediction is given by majority decision or by averaging past hypotheses. Figures 6 and 7 show a schematic diagram of QbagS and its pseudo-code, respectively.† It should be noted in Fig. 7 that the double for loop in Fig. 1 is now a single loop. By this modification, the

---

†C4.5 was used with the default parameters in all experiments described in this paper.

†In Fig. 7, a discrete-valued label is assumed.
number of points selected at each trial can be greatly increased, and it is likely that the computation efficiency will be greatly improved.

According to the survey paper of Provost and Kolluri [16] on mining techniques for large-scale data, QbagS is assigned to the category called “sequential multisubset learning with model-guided instance selection.” A recent method which belongs to this category and which apparently has the highest accuracy is Ivotes [4]. The authors performed a performance evaluation of QbagS in comparison to Ivotes. The flow of processing in Ivotes is the same as that in QbagS, but there is a decisive difference in the selection of the input data for the learning algorithm at the lower level. In QbagS, data labels are not utilized, and the points for which the prediction is the most scattered are selected. Ivotes, on the other hand, utilizes data labels, and the points for which prediction fails, that is, errors, are used as the input data.

Churn analysis of the above data was performed using both QbagS and Ivotes. QbagS and Ivotes were then compared on the basis of the learning curve for accuracy, and the relation between precision and recall after sufficient learning was investigated, as shown in Figs. 8 and 9. Recall refers to the percentage of correct prediction of the actual cancelers, and is in a complementary relation to the precision. In Fig. 8, C4.5 is used as the learning algorithm at the lower level in both QbagS and Ivotes. In Fig. 9, the CART decision tree [5] is used. As in the previously described experiment, five-fold cross-validation was applied.

Table 2 summarizes the results of Figs. 8(a) and 9(a). In the table, the computation time of QbagS is the time until the accuracy reaches the average accuracy of Ivotes for the final 10 trials. Z is the comparison test index of Weiss and Indurkhy [22]. It is calculated as follows, and the result is considered as statistically significant if Z is larger than 2:

\[
Z = \frac{\text{average accuracy of method A} - \text{average accuracy of method B}}{\sqrt{\text{variance of accuracy of method A} + \text{variance of accuracy of method B}}}
\]

We see from the results shown in Figs. 8(a) and 9(a) and Table 2 that QbagS provides higher accuracy than Ivotes for any learning algorithm at the lower level, and also has an advantage from the viewpoint of computational efficiency.

Algorithm: Query-by-Bagging:Single (QBagS)

**Input:** number of trials: M; learning algorithm at lower level: \( A \)
- number of candidate enquiry points: \( R \); number of selected points: \( D \); set of labels: \( Y \).
- candidate for query point in \( i \)-th trial: \( C_i \)
- selected point in \( i \)-th trial: \( S_i \)

**Initialization**
1. initial sample \( S_1 = (x_1, y_1, \ldots, x_D, y_D) \) is selected at random from database.
2. \( S_1 \) is used as the input to operate \( A \), and hypothesis \( h_1 \) is obtained.

**For** \( i = 1, \ldots, M \)
1. \( R \) of \( C_i \) are selected at random from database.
2. **For** all \( x \in C_i \), the following margin \( m(x) \) is calculated on the basis of past hypotheses \( h_1, \ldots, h_i \), respectively.

\[
m(x) = \max(1(x) - \max(2(x))
\]
where
\[
\max(1(x)) = \max_{t \leq i} \{ h_t(x) = y \}
\]
\[
\max(2(x)) = \max_{y \neq y} \{ t \leq i : h_t(x) = y \}
\]
\[
y_{\max}(x) = \arg \max_j \{ t \leq i : h_t(x) = y \}.
\]
3. \( D \) points \((x_1^*, y_1^*), \ldots, (x_D^*, y_D^*)\) with the smallest \( m(x) (x \in C_i) \) are selected from \( C_i \), and we set \( S_{i+1} = (x_1^*, y_1^*), \ldots, (x_D^*, y_D^*) \).
4. Using \( S_{i+1} \) as the input to operate \( A \), hypothesis \( h_{i+1} \) is obtained.

**End For**

**Output:** The final hypothesis is defined as follows:

\[
h_{\text{final}}(x) = \arg \max_{y \in Y} \{ t \leq M : h_t(x) = y \}
\]

Fig. 7. Query by bagging:single loop (QbagS).
Fig. 8. (a) Learning curves and (b) precision–recall curves of QbagS and Ivotes (component algorithm:C4.5).

Fig. 9. (a) Learning curves and (b) precision–recall curves of QbagS and Ivotes (component algorithm:CART).

Table 2. Comparison summary of QbagS and Ivotes

<table>
<thead>
<tr>
<th>Learning algorithm at lower level</th>
<th>QbagS</th>
<th>Ivotes</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>Computation time (s)</td>
<td>Accuracy (%)</td>
</tr>
<tr>
<td>C4.5</td>
<td>90.40</td>
<td>1207.3(1.0)</td>
<td>89.43</td>
</tr>
<tr>
<td>CART</td>
<td>90.80</td>
<td>1857.7(1.0)</td>
<td>90.28</td>
</tr>
</tbody>
</table>
QbagS, with a difference as great as 0.25. Thus, when we attempt to make the prediction correct for the top 10% of the actual cancelers, the accuracy between the two is as great as 25%. In business activity based on CRM, it is highly desirable to achieve highly accurate customer segmentation and customer targeting, and QbagS seems very well suited to this purpose.

The results shown in Figs. 8 and 9 are considered as typical cases in which the principle of maximizing the information content is better than the error-driven principle, such as that used in Ivotes, in selective sampling from actual large-scale data with a large amount of noise.

6. Conclusions

The similarity between ensemble learning and mass enquiry learning is noted, and a new set of procedures proposed by the authors, which can be called “active ensemble learning,” is described. The effectiveness of the approach has already been demonstrated by other related studies following the authors’ publications [11, 15, 18], and the approach is making a stream in machine learning and data mining.

Acknowledgments. The authors thank Mr. Goto, Manager at NEC, Mr. Sakata, President of Internet Systems Research Laboratories, and Mr. Dohi, Integration Manager of NEC Basic Research Laboratories, for the opportunity to perform this research. They also thank Professor Udaka of Kyoto University for his assistance in the application of active learning experiment planning to the MHC binding peptide identification problem, and for providing them with useful knowledge.

REFERENCES

Hiroshi Mamitsuka received his B.S. degree from the Department of Biochemistry, University of Tokyo, in 1988, completed the M.E. program (information engineering) in 1991, and joined NEC. Since 2002, he has been a visiting associate professor at the Chemical Research Laboratory, Kyoto University. He holds a D.Sc. degree.

Naoki Abe completed the M.E. program (information science) at MIT in 1984 and joined the IBM Thomas Watson Laboratories for one year. He completed the doctoral program (information science) at the University of Pennsylvania in 1989. He was a researcher at the University of California, Santa Cruz, for one year beginning in 1989. He joined NEC in 1990, and since 2001 has been affiliated with the IBM Thomas Watson Laboratories. He holds a Ph.D. degree.