A Theoretical Framework on the Ideal Number of Classifiers for Online Ensembles in Data Streams

Hamed R. Bonab*
Bilkent Information Retrieval Group
Computer Engineering Department
Bilkent University, 06800, Ankara, Turkey
hamed@bilkent.edu.tr

Fazli Can
Bilkent Information Retrieval Group
Computer Engineering Department
Bilkent University, 06800, Ankara, Turkey
canf@cs.bilkent.edu.tr

ABSTRACT

A priori determining the ideal number of component classifiers of an ensemble is an important problem. The volume and velocity of big data streams make this even more crucial in terms of prediction accuracies and resource requirements. There is a limited number of studies addressing this problem for batch mode and none for online environments. Our theoretical framework shows that using the same number of independent component classifiers as class labels gives the highest accuracy. We prove the existence of an ideal number of classifiers for an ensemble, using the weighted majority voting aggregation rule. In our experiments, we use two state-of-the-art online ensemble classifiers with six synthetic and six real-world data streams. The violation of providing independent component classifiers for our theoretical framework makes determining the exact ideal number of classifiers nearly impossible. We suggest upper bounds for the number of classifiers that gives the highest accuracy. An important implication of our study is that comparing online ensemble classifiers should be done based on these ideal values, since comparing based on a fixed number of classifiers can be misleading.

Keywords

Big data stream; ensemble size; weighted majority voting

1. INTRODUCTION

Online ensembles for data stream classification has gained a great importance over the past few years in big data research. Extensive empirical results show that combining a suitable number of classifiers improves the accuracy of predictions, versus a single classifier [3, 4, 5, 14]. However, increasing the number of component classifiers of an ensemble results in greater computational resource requirements in terms of time and memory. The high volume and velocity of big data streams make this even more crucial in online environments.

There is a limited number of studies for batch mode ensembles. Latinne et al. [12] propose a simple empirical procedure for limiting the number of classifiers based on the McNemar non-parametric test of significance. Similar approaches [2, 8], suggest a range of 10 to 20 base classifiers for bagging depending on its particular base classifier and dataset. Oshiro et al. [13] cast the idea that, there is an ideal number of component classifiers within an ensemble, which exploiting more base classifiers would bring no significant performance gain and would only increase computational costs, using the weighted average area under the ROC curve (AUC) and some dataset density metrics. Fumera et al. [7, 8] apply an existing analytical framework for the analysis of linearly combined classifiers of bagging, using misclassification probability. Hernández-Lobato et al. [10] suggest a statistical algorithm for determining the size of an ensemble, by estimating required number of classifiers for obtaining stable aggregated predictions, using majority voting.

The contributions of our work are the following: (a) To the best of our knowledge, we are first to study the number of component classifiers of online ensembles using weighted majority voting; (b) We theoretically model online ensembles for data streams using a geometric framework and prove that for the highest prediction accuracy, the number of classifiers should be the same as the number of class labels; (c) We experimentally examine our hypothesis, using two state-of-the-art ensemble classifiers and several synthetic and real-world data streams. We show the existence of the ideal number of classifiers; however, the violation of providing independent component classifiers makes exactly determining the ideal number of classifiers nearly impossible; and (d) We suggest an upper bound for the number of classifiers that gives the highest accuracy. Our study also shows that comparing online ensemble classifiers should be done based on these ideal values, as comparing based on a fixed number of classifiers can be misleading.

2. A GEOMETRIC FRAMEWORK

Motivation. We propose a framework for studying the theoretical side of online ensemble classifiers over data streams, based on [6]. Its general schema is presented in Fig. 1. Individual scores of component classifiers are modeled in a spatial environment as vectors for establishing a relationship between geometric features of vectors, and their corresponding effectiveness. Euclidean norm is used as the loss function for optimization purposes. According to [9], there are clear

*Present e-mail address: hamed.bonab@stonybrook.edu

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983907

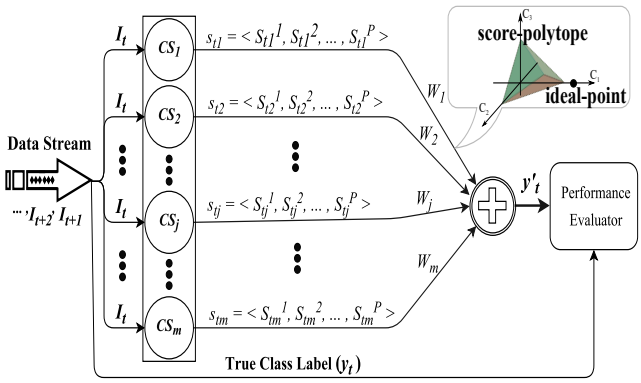


Figure 1: Schema of our geometric framework.

advantages of using the Euclidean norm for the LSQ solution. For aggregation, we use the weighted majority voting rule.

Optimum Weight Assignment. As shown in Fig. 1, we have an ensemble of component classifiers $\xi = \{CS_1, CS_2, \cdots, CS_m\}$. We assume that all $CS_j (1 \leq j \leq m)$ are independent of each other. For continual updates of the weights, we use the n latest instances $I = \{I_1, I_2, \cdots, I_n\}$, as an instance window, where I_n is the latest instance and all of the true class labels are available, $C = \{C_1, C_2, \cdots, C_p\}$. For each instance $I_i (1 \leq i \leq n)$, each component classifier $CS_j (1 \leq j \leq m)$ has a score-vector of $s_{ij} =< S_{ij}^1, S_{ij}^2, \cdots, S_{ij}^p >$. Mapping all score-vectors of I_i into a p-dimensional space, builds a polytope which we call the *score-polytope* of I_i. For the true class label of I_i we have $o_i =< O_i^1, O_i^2, \cdots, O_i^p >$ as the ideal-point. We aim to find the optimum weight vector $w =< W_1, W_2, \cdots, W_m >$, to minimize the distance between the score-polytope and ideal-point. Using the squared Euclidean norm as our measure of closeness for the linear least squares problem (LSQ) results

$$\min_w ||o - Sw||_2^2 \tag{1}$$

Where for each instance I_i, $S \in \mathbb{R}^{m \times p}$ is the matrix with relevance scores s_{ij} in each row, w is the vector of weights to be determined, and o is the vector of the ideal-point [9]. Since we have n instances in our window, we use the following function for our optimization solution.

$$f(W_1, W_2, \cdots, W_m) = \sum_{i=1}^{n} \sum_{k=1}^{p} (\sum_{j=1}^{m} (W_j S_{ij}^k) - O_i^k)^2 \tag{2}$$

Taking a partial derivation over $W_q (1 \leq q \leq m)$, setting the gradient to zero $\nabla f = 0$, and finding optimum points give us the optimum weight vector. Letting the following summations as a_{qj} and d_q

$$a_{qj} = \sum_{i=1}^{n} \sum_{k=1}^{p} S_{iq}^k S_{ij}^k, \quad (1 \leq q, j \leq m) \tag{3}$$

$$d_q = \sum_{i=1}^{n} \sum_{k=1}^{p} O_i^k S_{iq}^k, \quad (1 \leq q \leq m) \tag{4}$$

lead to m linear equations with m variables (weights). The proper weights in the following matrix equation are our intended optimum weight vector. Briefly, $Aw = d$, where A is the coefficients matrix and d is the remainder vector.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \times \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_m \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{bmatrix} \tag{5}$$

Discussion. According to Eq. 3, A is a symmetric square matrix. If A has full rank, our problem has a unique solution. On the other hand, in the sense of a least squares solution [9], it is probable that A is rank-deficient, and we may not have a unique solution. Studying the properties of this matrix lead us to the following theorem.

THEOREM 1. *If the number of component classifiers is not equal to the number of class labels, $m \neq p$, then the coefficient matrix would be rank-deficient,* $\det A = 0$.

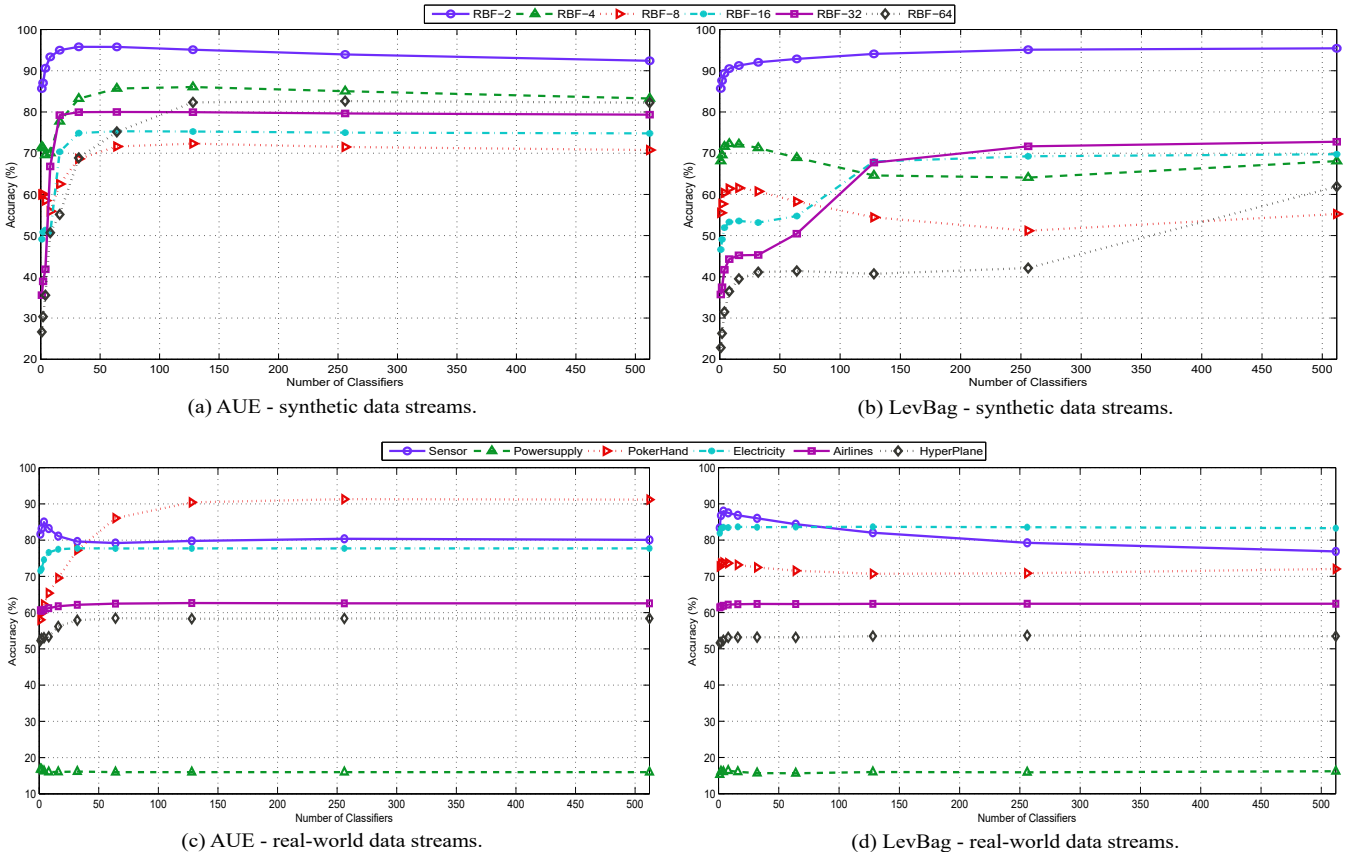
PROOF. Since we have p dimensions in our Euclidean space, p independent score vectors would be needed for the basis spanning set. Any number of vectors, m, more than p is dependent on the basis spanning set, and any number of vectors, m, less than p is insufficient for constructing the basis spanning set. □

The above theorem excludes some cases in which we cannot find optimum weights for aggregating votes. In addition, for cases where $m = p$, letting $\det A = 0$ for the parametric coefficient matrix results in some conditions that we have vote agreement, and cannot find a unique optimum weight vector. As an example, suppose that we have 2 component classifiers for a binary classification task, $m = p = 2$. Letting $\det A = 0$, will result in following equations; $S_{11}^1 + S_{12}^2 = 1$ or $S_{11}^2 + S_{12}^1 = 1$, meaning the agreement of component classifiers. This gives better insight for the commonly agreed upon idea that "the classifiers should be different from each other, otherwise the overall decision will not be better than the individual decisions" [11]. Optimum weights can be reached when we have the same number of independent and diverse component classifiers as class labels.

Conclusion. This theorem supports the idea that there is an ideal number of component classifiers for an ensemble, with which we can reach the most accurate results. Increasing or decreasing the number of classifiers from this ideal point would deteriorate predictions. We refer to this situation as "the law of diminishing returns in ensemble construction." Our framework suggests the number of class labels of a dataset as the ideal number of component classifiers, with the premise that they generate independent scores. However, real-world datasets and existing ensemble classifiers do not guarantee this premise most of the time. Determining the exact value of this ideal point for a given ensemble classifier, over a real-world data stream, is still a challenging problem.

3. EXPERIMENTAL SETUP

In our experiments, we show the relationship between the number of classifiers and number of class labels for 12 different data streams and 2 online ensembles. We conduct our experiments on Accuracy Updated Ensemble (AUE)[5] and Leverage Bagging (LevBag)[3] ensembles using synthetic and real-world data streams. For implementation, we used the MOA framework; for evaluation, we used the Interleaved Test-Then-Train approach [4].



(a) AUE - synthetic data streams.

(b) LevBag - synthetic data streams.

(c) AUE - real-world data streams.

(d) LevBag - real-world data streams.

Figure 2: Prediction behavior of AUE and LevBag ensembles, in terms of accuracy, with different number of classifiers ranging from 2^0 to 2^9 by doubling at each step, for synthetic and real-world data streams.

Table 1: Summary of Real-world Data Streams

Dataset	#Instances	#Attr.	#Class Labels
Sensor [1]	2,219,803	5	54
Powersupply [1]	29,928	2	24
PokerHand [4]	1×10^7	10	10
Electricity [4]	45,312	8	2
Airlines [4]	539,383	7	2
HyperPlane [1]	1×10^7	10	5

3.1 Data Streams

There are a few real-world data streams in the literature. Finding proper data streams, where the number of class labels is the only variant, is nearly impossible.

Synthetic Data Streams. We chose the popular Random RBF generator, since it is capable of generating data streams with an arbitrary number of features and class labels [4]. Using this generator implemented in the MOA framework, we prepared 6 different datasets, each containing 1 million instances with 20 attributes. The number of class labels is chosen as 2, 4, 8, 16, 32, and 64. We reflect these in the naming of our RBF data streams (RBF-2, RBF-4, etc.).

Real-world Data Streams. We selected 6 different real-world datasets used as data streams in the literature—they all are available online for downloading [1, 4]. Table 1 gives a summary of these data streams.

3.2 Ensemble Methods

We used 2 state-of-the-art ensemble classifiers: AUE and LevBag that use the weighted majority voting and majority voting aggregation rules, respectively, and have different component classifier training strategies. We exploited Hoeffding trees as base component classifiers for both ensembles [4]. Ensemble sizes in terms of the number of base classifiers grow exponentially from 2^0 to 2^9, by doubling at each step. This range is chosen to show the behavior in a large-scale of possibilities.

4. RESULTS AND DISCUSSION

Observations. Fig. 2 shows the impact of doubling the ensemble sizes on the accuracy of predictions. We plot the experimental results with the real numerical scale rather than the logarithmic, since it provides a better exposition of the results. In each plot, the existence of a peak point in which an online ensemble reaches its maximum accuracy is well-illustrated. For the ever-increasing cases, we believe that there is a peak point out of our range. Computational requirements, in term of CPU time and memory usage, are roughly doubled with doubling ensemble size.

Following our theoretical expectations, AUE provides higher consistency compared to LevBag, as seen in Fig. 2-(a) vs. (b) and (c) vs. (d). Since synthetic data streams have a large number of attributes, 20, the possibility of training independent and diverse component classifiers is high [11,

Table 2: Highest/Peak Accuracy Values in Percentage and the Corresponding Number of Component Classifiers of Ensembles for Different Data Streams

	AUE		LevBag	
Dataset	Acc.	#Classifiers	Acc.	#Classifiers
RBF-2	95.80	32	95.44	512
RBF-4	86.05	128	72.27	8
RBF-8	72.33	128	61.58	16
RBF-16	75.33	64	69.76	512
RBF-32	79.80	64	72.77	512
RBF-64	82.64	256	61.91	512
Sensor	84.99	4	87.99	4
Powersupply	16.74	1	16.34	8
PokerHand	91.30	256	73.82	4
Electricity	77.73	32	83.66	16
Airlines	62.63	128	62.43	256
HyperPlane	58.47	64	53.68	256

13]. For example, Fig. 2-(a) shows that for a lower number of class labels (RBF-2, -4, and -8) the results are highly consistent with our theoretical expectations. On the other hand, for larger number of class labels (RBF-16, -32, and -64) we observe lower consistency. The impact of the number of attributes and class labels is also true for real-world data streams: PokerHand, HyperPlane, and Electricity provides higher consistency than those of Sensor, Powersupply, and Airlines.

Statistical Analysis. Table 2 shows the highest/peak accuracy values and corresponding number of classifiers for both ensembles. We conducted the Wilcoxon signed-ranks test for further analysis of our results, with $\alpha = 0.05$. Comparing the accuracy peaks of AUE and LevBag results in 10 positive and 2 negative differences. The two-tailed probability value, $P = 0.027$, is less than 0.05; it can be accepted that the accuracy peaks of AUE are statistically significantly higher than those of LevBag. Comparing the number of component classifiers on these peak values results in 4 positive and 7 negative differences, with two-tailed probability value, $P = 0.147$. While there is no statistically significant difference between the numbers of component classifiers for the two algorithms, AUE reaches higher peaks compared to LevBag. Hence, comparing ensembles based on a fixed number of component classifiers can be misleading. However, studies based on a fixed number of classifiers can be acceptable since in such cases all ensemble methods can be equally disadvantaged. For example, without ensuring fairness, comparing the accuracy of AUE and LevBag ensembles each with 8 or 16 base classifiers, which are in the range of conventional comparisons [5, 14], results in no statistically significant difference in our experiments.

We compared the observed peak accuracy values of both ensembles using the non-parametric Friedman statistical test with $\alpha = 0.05$ and $F(2, 22)$. The resulting two-tailed probability value, $P = 0.032$, rejects the null-hypothesis. The multiple comparisons show that the theoretical number of classifiers are statistically significantly different from the practical peak accuracy values of both algorithms. However, multiplying these theoretical number of classifiers by a constant value, in our case 2, makes the differences statistically insignificant. This can be used for obtaining upper-bounds of the ideal number of classifiers for a given data stream and ensemble classifier.

5. CONCLUSION AND FUTURE WORK

Our model showed that: (1) Theoretically, using the same number of independent component classifiers as class labels gives the highest prediction accuracy; (2) Practically, due to the violation of independency of component classifiers, determining these peak values is nearly impossible. However, upper bounds can be considered for this problem and that needs further investigation. An important implication of our study is that comparing online ensemble classifiers should be done based on these peak values, since comparing based on a fixed number of classifiers can be misleading.

6. ACKNOWLEDGEMENTS

We appreciate Jon M. Patton from Miami Univ. of OH, and Alper Can for their valuable comments on this work.

7. REFERENCES

[1] X. Zhu, Stream Data Mining Repository, http://www.cse.fau.edu/ xqzhu/stream.html, 2010.

[2] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn.*, 36(1-2):105–139, 1999.

[3] A. Bifet, G. Holmes, and B. Pfahringer. Leveraging bagging for evolving data streams. In *ECML PKDD*, pages 135–150, 2010.

[4] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavaldà. New ensemble methods for evolving data streams. In *ACM SIGKDD*, pages 139–148, 2009.

[5] D. Brzezinski and J. Stefanowski. Reacting to different types of concept drift: The accuracy updated ensemble algorithm. *IEEE TNNLS*, 25(1):81–94, 2014.

[6] L.-W. Chan. Weighted least square ensemble networks. In *IJCNN*, volume 2, pages 1393–1396, 1999.

[7] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE TPAMI*, 27(6):942–956, 2005.

[8] G. Fumera, F. Roli, and A. Serrau. A theoretical analysis of bagging as a linear combination of classifiers. *IEEE TPAMI*, 30(7):1293–1299, 2008.

[9] P. C. Hansen, V. Pereyra, and G. Scherer. *Least Squares Data Fitting with Applications*. JHU Press, 2013.

[10] D. Hernández-Lobato, G. Martínez-Muñoz, and A. Suárez. How large should ensembles of classifiers be? *Patt. Recog.*, 46(5):1323–1336, 2013.

[11] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, 2003.

[12] P. Latinne, O. Debeir, and C. Decaestecker. Limiting the number of trees in random forests. In *MCS*, pages 178–187, 2001.

[13] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas. How many trees in a random forest? In *MLDM*, pages 154–168, 2012.

[14] N. C. Oza and S. J. Russell. Experimental comparisons of online and batch versions of bagging and boosting. In *ACM SIGKDD*, pages 359–364, 2001.