Active Learning Strategy for Text Categorization Based on Support Vectors Relative Positioning

Vladimir Vakurin[1], Andrey Kopylov[1], Oleg Seredin[1], and
Konstantin Mertsalov[2]

[1] Tula State University, Russia
vakourinvl@yandex.ru, and.kopylov@gmail.com, oseredin@yandex.ru,
[2] Rensselaer Polytechnic Institute, USA
kmertsalov@gmail.com

Abstract. A method of decreasing the number of requests to a human labeler, required for annotation of a large text corpus, is proposed. We use an active learning strategy based on subdivision of labeling process into iterative steps, starting from some initial training set and using SVM classification results to select a set of objects to be labeled by an expert and added to the training set on the next step. Such procedure can significantly reduce time an amount of objects needed for classifier training without loss of recognition accuracy.

Key words: Active Learning, SVM margin, Text Categorization.

1 Introduction

We address an issue of an efficient use of time of human reviewers that are often employed to review and categorize electronic texts to create training sets for the Supervised and Semi-supervised learning methods. Applications of such methods aim to use the training data to develop an automated classifiers capable of annotating arbitrarily large data sets. However, the construction of such classifier is constrained by the time and cost required to review and categorize sufficiently large training set by human reviewers. The research problem addressed here stems from the trade off between the need to develop a large enough training set required for an accurate classifier and the need to control the costs of creating such training set which limits the number of documents that can be reviewed by humans.

We consider a method of choosing text objects to be reviewed by the human reviewer from the pool of available data in a way that accelerates the learning process. The proposed method belongs to the group of online recognition methods, and is based on the analysis of the interclass border areas of the general assembly. Experimental results show the effectiveness of the method in comparison with other methods of Active Learning in spite of its relative simplicity.

2 Related Work

All Machine learning approaches can be divided into two large groups, namely online [24] and offline [3] methods, by the method they use to obtain and process training data.

In case of offline approaches, it is assumed that the whole training set is available for analysis and remains unchanged during the functioning of a recognition system, while in online systems a new labeled objects or even a sets of labeled objects, unavailable at the initial training stage, become available.

If the data, used for training, is no longer available or limited and can not be effectively used, when all the information about this part of data is represented only through a decisive rule. Machine learning approaches, which deal with such situation are referred to as Incremental Learning [1,11], or, like in the early Soviet works, Recurrent Learning. Sometimes Incremental Learning is used as a synonym to Online Learning.

Reinforcement Learning [29] is an area of machine learning in which the considered system or agent trains in the process of interaction with some environment. Reinforcement signals are formed by the reaction of the environment to accepted decisions but not the special reinforcement control system like in the case of supervised learning. Therefore reinforcement learning is a particular case of supervised learning, but the environment itself or its model plays the role of a teacher. It is also necessary to take into account that some reinforcement rules are based on implicit teachers, for example, in the case of an artificial neural network, on the simultaneous activity of formal neurons, so that they can be attributed to unsupervised learning [13].

A situation, similar to online recognition, can occur when the training set is so large that the available computational tools do not allow to process it entirely. Such a situation was typical for the initial stages of the development of the theory of machine learning, when computing resources were extremely limited. Now this problem again comes to the fore in relation with the increased amount of available data produced by the informational systems and gives rise to a new area of research known as "Big Data".

Some pattern recognition tasks have a specificity, that makes it possible to separate them into a class of problems with semi-labeled sample set (Semi-supervised Learning) [23,35,36], when the features of other objects but not classification labels is known in additional to the training set. However, the presence of such objects gives the necessary information about the population. This additional information can be used to accelerate the training process or to increase accuracy. The tasks, when we can request a class label for some objects from the teacher, originate an important subclass of such problems. Object selection strategies with the correspondent methods of correction of decision rules form the Active Learning subclass of pattern recognition problems [26].

If there are data about an object's class received from several experts, or the accuracy of the teacher can be questioned and has some degree of reliability, when the further synthesis of active learning methods, called Proactive Learning can be maid [10].

A comprehensive survey, devoted to the problems of informational retrieval on the text data, ways of arrangement of the experiments and evaluation of their results can be found in the article [25]. Problems of generalization ability improvement and active learning tasks are considered in [26,27,21,17].

This paper is attributed to a known pool-based learning problem [2] of labeling objects from some subset in general assembly, and further model retraining to improve its generalization ability on that population.

A technique of queries the objects about which it is least certain how to label for further expert annotation (uncertainty sampling) is described in [9] and, in application to support vector machines, the object selection from interclass margin in [31]. The common characteristic feature of such techniques consists in using some classification uncertainty measure so that the borderline objects are the most preferable for disclosure by an expert [31]. For example, when using a probabilistic model for binary classification, uncertainty sampling simply queries the instance whose posterior probability of being positive is nearest 0.5. For linear classifiers (like linear SVM) such objects are those that are close to the margin between classes.

The set of hypotheses consistent with the current labeled training set is called the version space (VS) [16]. The generalization error of most possible classifiers of the population in the VS (named by authors - egalitarian generalisation error bound) is controlled by the size of the VS relative to the size of the hypothesis space [5]. Learner that chooses successive queries that halves the version spaces is the learner that minimizes the maximum expected size of the version space (p.51 in [31]). If the object is placed near the hyperplane the rate of dimensionality reduction approaching two in accordance with the distance from this object to the hyperplane and does not depends from real class label. This rule holds true if the training set can be linearly separated (though this requirement is not considered too strict by the author (p.49 in [31]).

We propose to use the distance from an object to support vectors as well as the distance to the hyperplane for object selection, as the compactness and symmetry assumptions is often not satisfied in practice [5] and the separating hyperplane could not intersect version space at all (p.52 in [31]). The trivial argument for such an object selection is the necessity to reduce the number of manually labeled duplicate texts.

According with [31] active learning means the directed strategy of choosing objects from general assembly for labeling. Good strategy allows to minimize number of queries to human labelers.

Several methods for choosing objects for class membership opening have investigated in [18,20]. Comparison of generalization performance increasing speedup for these two techniques were published in[15]. In [30] the demo program have introduced for active learning algorithm based on Naive Bayes classifier, generalization performance increasing speedup and users survey.

In [28] expert provides explanations and produces labeling of domain featured fragments of texts. Based on these fragments of interests and some other specify fragments which oppose the whole document label authors introduce method of

Learning with Explanations. In [34] the problem of active learning for networked data, where samples are connected with links and their labels are correlated with each other have studied. Authors particularly focus on the setting of using the probabilistic graphical model to simulate the networked data, due to its effectiveness in capturing the dependency between labels of linked samples. Two-stage active learning technique for multi-label problem were suggested in [8] and summarize principles from [7]. In first stage an effective multi-label classification model by combining label ranking with threshold learning, which is incrementally trained to avoid retraining from scratch after every query was introduced. Then based on this model, authors propose to exploit both uncertainty and diversity in the instance space as well as the label space, and actively query the instance-label pairs which can improve the classification model most.

It is noticed from [31] that there is no any reason to prefer for labeling one of the classes. So, we base our strategy on principle of equal significance for label opening for class of interest and other.

3 Active Learning Algorithm based on Support Vectors Relative Positioning

We are using the principle of uncertainty sampling and analysis of borderline objects. Corrections of decision rule implemented via opening labels for objects which are close to decision boundary. It is well-known fact that support vectors in Vapniks SVM [32] uniquely define the so called separated hyperplane. The score function (according to sign) for any unlabeled test object (document, text) will be more reliable to class membership. The our idea of suggested algorithm is to take for testing by human labeler just such documents which are close to the decision boundary and at the same time are far from set of support objects. So the formal algorithm is follows:

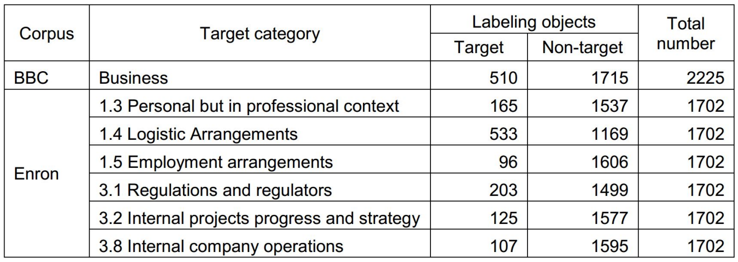
1. At the first step the optimal separating hyperplane is built for initial training set and the subset of support objects is fixed. Initial training set can be obtained via review of documents sampled randomly from the complete population.
2. For all unlabeled objects take into account just those H which is closest to hyperplane.
3. From H for labeling it is necessary to select N objects $\omega_i \in H$ under condition $L_i > L_j,\ i = 1, .., N,\ j > N$, where $L_j = \min\limits_{\omega_k^{SV}} \left(d(\omega_j, \omega_k^{SV}) \right)$ - is the distance from object $\omega_j \in H$ to the set of support objects ω_k^{SV}.

The distance is a regular Euclidean distance in corresponding feature space.

4 Experimental Study. The Procedure

It is clear that computational costs for each iteration of such experiment are considerable and rise the problem of time sharing between computer (decision

Table 1. Data set characteristics



Corpus	Target category	Labeling objects		Total number
		Target	Non-target	
BBC	Business	510	1715	2225
Enron	1.3 Personal but in professional context	165	1537	1702
	1.4 Logistic Arrangements	533	1169	1702
	1.5 Employment arrangements	96	1606	1702
	3.1 Regulations and regulators	203	1499	1702
	3.2 Internal projects progress and strategy	125	1577	1702
	3.8 Internal company operations	107	1595	1702

rule correction) and human labeler (reading documents and take a decision about categorization). Moreover the portion of texts presented for one iteration should be large enough to be representative but not too large to give the expert an a reasonable opportunity to read and consider each document before labeling. For our experiments we have chosen 20 documents for labeling which is consistent with discussions published in (p.49 in [19]). For the experimental study we used labeled data set (general assembly imitation) which was divided on three parts: **initial training (start) set, selection set** and **verificatory set**. These three sets are defined randomly at the beginning of each experiment in proportion of 5%, 45% and 50% accordingly. For all competitor methods the start and verificatory sets were fixed and isolated from selection set. So, we provide the objective monitoring of decision rule quality and generalization performance increasing speedup while transpose 20 objects from selection set into training set according to active learning strategy.

In the case when our algorithm does not take 20 objects ($N < 20$, underrun of objects in the $\omega_j \in H$) we take missing objects by random choice.

We use two-class SVM with linear kernel as routine classifier. The value of parameter C was equal to 100000. The feature space were formed with frequency properties of text documents using TF*IDF technique [22], the number of features were limited by 15 000 of most high rate in a corpus.

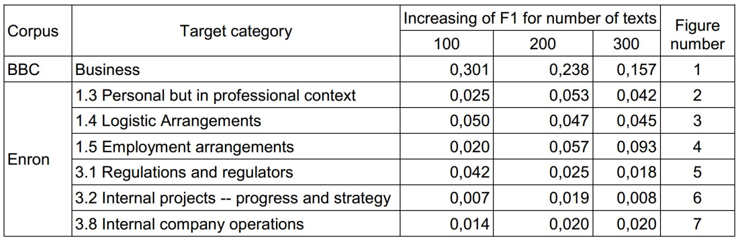
The core of our algorithm has the same essence as [14] but we have no idea to try to find the representative examples. From our point of view it is not possible. Moreover we apply our method to the task of text classification not for the well known databases.

5 Experimental Study. Results

Two corpuses of texts were used for experimental study: Enron with categories [12], BBC news [4]. Statistics for these datasets are presented in in Table 1.

The quality of active learning strategy were evaluated by common indices: average values of precision, recall, F1 [6] for multiple splitting data on training initial training set, selection set and verificatory set. For the most of published results number of splitting was equal to 100. So the total computation costs were near two months of Intel Core i5 4*4GHz 4GB.

Table 2. Quality inprovment



Corpus	Target category	Increasing of F1 for number of texts			Figure number
		100	200	300	
BBC	Business	0,301	0,238	0,157	1
Enron	1.3 Personal but in professional context	0,025	0,053	0,042	2
	1.4 Logistic Arrangements	0,050	0,047	0,045	3
	1.5 Employment arrangements	0,020	0,057	0,093	4
	3.1 Regulations and regulators	0,042	0,025	0,018	5
	3.2 Internal projects -- progress and strategy	0,007	0,019	0,008	6
	3.8 Internal company operations	0,014	0,020	0,020	7

The formal quality estimation for the three stages of each experiment - reading of expert (automatic opening of labels and moving objects into the training set) for 100, 200 and 300 documents from selective part (see Table 2). The charts (Figures 1 and 2) demonstrated the increasing of average quality for two methods - our (the blue line) and random choice of 20 objects (red line).

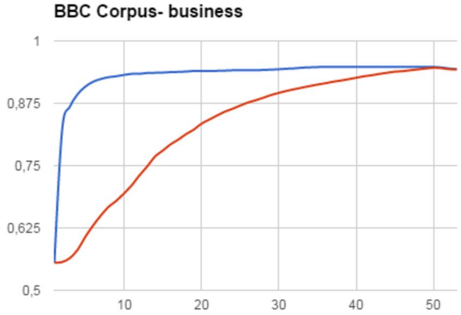
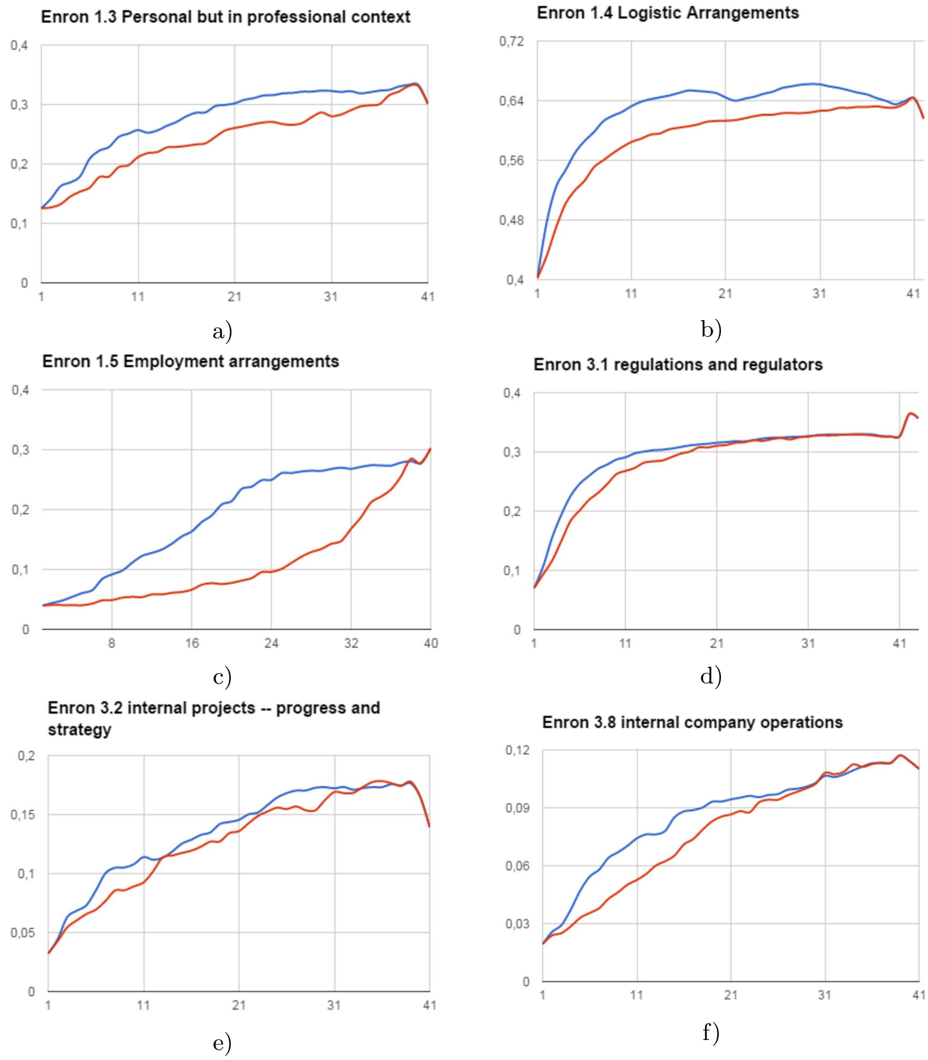


Fig. 1. Comparing curves of quality on the test set for BBC dataset. Average quality (F1) growth throw steps of algorithm by 20 documents. The blue line - our method, red one - random choice.

For the comparative estimation of quality improving of our method we use as competitor the algorithm of multi-class classification with active learning [8]. Results are published for Enron corpus, see figure 1 from named above paper. Note, that decision of the classifier assumed to be correct when it matches the decision of at least one expert. Due to another rising of task (multi-class classification) in the Active learning based on Uncertainty and Diversity for Incremental multi-label learning , AUDI [8] parameter micro F1 [33] is used.

Results are presented in the Table 3. It is obvious that algorithms show comparable growth of quality.



a)

b)

c)

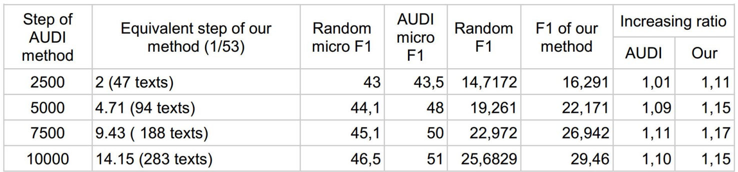
d)

e)

f)

Fig. 2. Comparing curves of quality on the test set for Enron dataset. Average quality (F1) growth throw steps of algorithm by 20 documents. The blue line - our method, red one - random choice.

Table 3. Results in comparing with AUDI method



Step of AUDI method	Equivalent step of our method (1/53)	Random micro F1	AUDI micro F1	Random F1	F1 of our method	Increasing ratio	
						AUDI	Our
2500	2 (47 texts)	43	43,5	14,7172	16,291	1,01	1,11
5000	4.71 (94 texts)	44,1	48	19,261	22,171	1,09	1,15
7500	9.43 (188 texts)	45,1	50	22,972	26,942	1,11	1,17
10000	14.15 (283 texts)	46,5	51	25,6829	29,46	1,10	1,15

6 Discussions and Conclusion

Experimental study demonstrates that suggested technique of active learning despite of its simplicity shows good results and can be used in industrial tasks. The quality improvement in comparing with random selection of objects on the BBC corpus was 23% and for different subsets of Enron database the was 3,4%. Reaching of the same quality in the test set our algorithm request in average 586 documents less than random choice for BBC corpus and 163 documents for Enron corpus. For our opinion the BBC corpus (news texts) used in a lot of investigations are not typical for industrial tasks, so the real improving of recognition quality will be closer to results on Enron with categories. Experimental study reveals the number of problems which will be interesting for future investigations:

1. which active learning algorithm appropriate to particular feature space;
2. how to use a priori information about non-labeled set in active learning;
3. what is the criterion that labeling process have reached quality saturation.

References

1. Angluin, D. Smith, C. A survey of inductive inference: Theory and methods// Computing Surveys. 1983. 15: P. 237-269
2. Cohn, D. A. Atlas, L. Ladner, R. E. Improving generalization with active learning // Machine Learning. 1994. 15(2). P. 201221.
3. Guillory A., Chastain E., Bilmes J. A. Active Learning as Non-Convex Optimization //AISTATS. 2009. p. 201-208.
4. Greene D., Cunningham P. Practical solutions to the problem of diagonal dominance in kernel document clustering //Proceedings of the 23rd international conference on Machine learning. ACM, 2006. p. 377-384.
5. Herbrich, R. Graepel, T. Williamson, R. C. The Structure of Version Space in Innovations in Machine Learning: Theory and Applications/ edited by Holmes D.E. , Jain L.C.// p. 263-279.-Berlin: Springer, (2) 2006.- ISBN 3-540-30609-9.
6. He, H. Learning from imbalanced data// IEEE Transactions on Knowledge and Data Engineering. (21.9). 2009. P. 1263-1284.
7. Huang S. J., Jin R., Zhou Z. H. Active learning by querying informative and representative examples //Advances in neural information processing systems. 2010. pp. 892-900.

8. Huang, S-J. Zhou, Z-H. Active Query Driven by Uncertainty and Diversity for Incremental Multi-Label Learning Data Mining/ (ICDM), IEEE 13th International Conference. 2013.- ISSN 1550-4786
9. Lewis, D. D. Catlett, J. Heterogeneous uncertainty sampling for supervised learning/ In Proceedings ICML 94, 1994. pages 148156.
10. Lin C. H., Mausam M., Weld D. S. Re-Active Learning: Active Learning with Relabeling //AAAI. 2016. p. 1845-1852.
11. Jantke P. Types of incremental learning //AAAI Symposium on Training Issues in Incremental Learning. 1993. p. 23-25.
12. Klimt, B. Yang, Y. The Enron Corpus: A New Dataset for Email Classification Research / in Proceedings ECML04. P. 217-226.- Pisa, Italy,2004.
13. Kohonen, T. The self organizing map /Proceedings of the Institute of Electrical and Electronics, vol. 78, P.14641480,1990.
14. Kremer J., Steenstrup Pedersen K., Igel C. Active learning with support vector machines //Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014. Vol. 4. No. 4. pp. 313-326.
15. Melville, P. Sindhwani,V. Active dual supervision: Reducing the cost of annotating examples and features //In Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. 2009. P. 4957.
16. Mitchell, T. Generalization as search // Artificial Intelligence,(28) 1982. P.203226.
17. Olsson, F. A literature survey of active machine learning in the context of natural language processing // SICS Report, T2009:06. 2009.-ISSN 1100-3154.
18. Fung G. M., Mangasarian O. L., Shavlik J. W. Knowledge-based support vector machine classifiers //Advances in neural information processing systems. 2003. p. 537-544.
19. Raghavan, H. Tandem learning: A learning framework for document categorization / Ph.D. thesis.-Amherst:Graduate school of Massachusetts, 2007.
20. Raghavan, H. Madani, O. Jones, R. Active Learning with Feedback on Both Features and Instances // Journal of Machine Learning Research. (7) 2006. P. 16551686.
21. Rubens, N. Elahi, M. Sugiyama, M. Kaplan, D. Active Learning in Recommender Systems //In ed. Ricci F., Rokach L., Shapira B., Recommender Systems Handbook (2 ed.).-US: Springer, 2016.- ISBN 978-1-4899-7637-6.
22. Salton, G. McGill, M. J. Introduction to modern information retrieval. McGraw-Hill., 1986.-ISBN 978-0070544840
23. Scudder, H. J. Probability of Error of Some Adaptive Pattern-Recognition Machines// IEEE Transaction on Information Theory, (11) 1965. P.363371. Cited in Chapelle et al. 2006, page 3.
24. Sculley D. Online active learning methods for fast label-efficient spam filtering //CEAS. 2007. Vol. 7. p. 143.
25. Sebastiani F. Machine learning in automated text categorization //ACM computing surveys (CSUR). 2002. Vol. 34. No. 1. p. 1-47.
26. Settles, B. Active Learning Literature Survey // Machine Learning. 2010. Vol. 15, No. 2. P. 201221.
27. Settles, B. Active Learning // Synth. Lectures Artificial Intelligence Mach. Learn. 2012. Vol. 6, No. 1. P. 1114.
28. Sharma, M. Bilgic, M. Towards Learning with Feature-Based Explanations for Document Classification.-IL:. Illinois Institute of Technology, Chicago, USA. 2016.
29. Sutton, R. S. Temporal Credit Assignment in Reinforcement Learning / (PhD thesis). -MA(US):.Amherst: University of Massachusetts,1984.

30. Stumpf S. et al. Integrating rich user feedback into intelligent user interfaces //Proceedings of the 13th international conference on Intelligent user interfaces. ACM, 2008. p. 50-59.
31. Tong, S. Koller, D. Support vector machine active learning with applications to text classification// JMLR The Journal of Machine Learning Research. (2) 2002. p. 45-66.
32. Vapnik V. Statistical Learning Theory. Wiley-Interscience. NY, 1998.
33. Yang Y. A study of thresholding strategies for text categorization //Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001. p. 137-145. doi: 10.1145/383952.383975 2.
34. Yang Z. et al. Active learning for networked data based on non-progressive diffusion model //Proceedings of the 7th ACM international conference on Web search and data mining. ACM, 2014. p. 363-372.
35. Zhu, Xiaojin Semi-supervised learning literature survey // Computer Sciences.-MA(WI,US):University of Wisconsin-Madison, 2008.
36. Zhu X., Goldberg A. B. Introduction to semi-supervised learning //Synthesis lectures on artificial intelligence and machine learning. 2009. Vol. 3. No. 1. p. 1-130.