Discretizing Continuous Attributes in AdaBoost for Text Categorization

by Pio Nardiello, Fabrizio Sebastiani, and Alessandro Sperduti

---

Researchers from the University of Padova and from ISTI-CNR, Pisa, are undertaking a collaborative effort aimed at producing better best text classification strategies through the design of methods for the discretization of continuous attributes. This will make it possible to exploit the rich information contained in the non-binary weights produced by standard statistical or probabilistic term weighting techniques, in the context of high performance learners (such as AdaBoost and its variants) requiring binary input.

In the last ten years an impressive array of learning techniques have been used in text categorization (TC) research. Among these, the two classes of methods that most seem to have caught the attention of TC researchers are boosting (a subclass of the classifier committees class) and support vector machines. The reasons for this attention are twofold, in the sense that both classes exhibit strong justifications in terms of computational learning theory and superior effectiveness once tested on TC benchmarks of realistic size and difficulty. It is on the former class of methods that our work focuses.

Classifier committees (aka ensembles) are based on the idea that, given a task that requires expert knowledge to perform, several experts may be better than one if their individual judgments are appropriately combined. In TC, this means applying several different classifiers to the same task of deciding whether a given document belongs or not to a given category, and then combining their outcome appropriately. Boosting is a method for generating a highly accurate classifier by combining a set of moderately accurate classifiers ('weak hypotheses'). In this work we make use of two algorithms, called AdaBoost.MH and AdaBoost.MH(KR), which are based on the notion of "adaptive boosting", a version of boosting in which members of the committee can be sequentially generated after learning from the classification mistakes of previously generated members of the same committee. AdaBoost.MH is a realization of the well-known AdaBoost algorithm, which is specifically aimed at multi-label TC (ie the TC task in which any number of categories may be assigned to each document), and which uses 'decision stumps' (ie decisions trees composed of a root and two leaves only) as weak hypotheses. AdaBoost.MH(KR) is a generalization of AdaBoost.MH previously designed and implemented by these authors (see ERCIM News 44, p. 55) and based on the idea of learning a committee of classifier sub-committees; in other words, the weak hypotheses of AdaBoost.MH(KR) are themselves committees of decision stumps. So far, both algorithms have been among the best performers in text categorization experiments run on standard benchmarks.

A problem in the use of both algorithms is that they require documents to be represented by binary vectors, indicating presence or absence of the terms in the document. As a consequence, these algorithms cannot take full advantage of the 'weighted' representations, consisting of vectors of continuous (ie non-binary) attributes that are customary in information retrieval tasks, and that provide a much more significant rendition of the document's content than binary representations. In this work we address the problem of exploiting the potential of weighted representations in the context of AdaBoost-like algorithms by discretizing the continuous attributes through the application of entropy-based discretization methods. These algorithms attempt to optimally split the interval on which these attributes range into a sequence of disjoint subintervals. This split engenders a new vector (binary) representation for documents, in which a binary term indicates that the original non-binary weight belongs or does not belong to a given sub-interval.

Although the discretization methods we present can also be used in connection with learners not belonging to the boosting family, we focus our experiments on AdaBoost.MH and AdaBoost.MH(KR). Our experimental results on the Reuters-21578 text categorization collection (the standard benchmark of TC research) show that for both algorithms the version with discretized continuous attributes outperforms the version with traditional binary representations. This improvement is especially significant since AdaBoost.MH and AdaBoost.MH(KR) are nowadays in the restricted lot of the peak text categorization performers, a lot where the margins for performance improvement are slimmer and slimmer.

Link:
http://faure.iei.pi.cnr.it/~fabrizio/Publications/ContAtt.pdf

Please contact:
Fabrizio Sebastiani, ISTI-CNR
Tel: +39 050 315 2892
E-mail: fabrizio@iei.pi.cnr.it