ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-771

How Many Objects?: Determining the Number of Clusters with a Skewed Distribution

Satoshi Oyama[1] and **Katsumi Tanaka**[2]

Abstract. We propose a supervised approach to enable accurate determination of the number of clusters in object identification. We use the aggregated attribute values of the data set to be clustered as explanatory variables in the prediction model. Attribute aggregation can be done in linear time with respect to the number of data items, so our method can be used to predict the number of clusters with a low computational burden. To deal with skewed target values, we introduce a two-stage method as well as a method using a higher-order combination of explanatory variables. Experiments demonstrate our methods enable more accurate prediction than existing methods.

1 INTRODUCTION

Object-identification problems, in which it is necessary to determine whether names appearing in documents or database records correspond to the same real world object, are important in information retrieval and integration. Typical examples of object-identification problems include disambiguating namesakes in Web search results and establishing correspondence between an abbreviated author name in bibliographic databases and a particular person. Object-identification problems are generally solved by clustering data that contain an ambiguous name and by regarding data in the same cluster as corresponding to the same object.

Among the various clustering algorithms, the most widely used are k-means and hierarchical algorithms including single-linkage. One problem in using the k-means clustering algorithm, though, is that a user has to specify the number of clusters as a parameter before starting the clustering procedure. If we use a hierarchical clustering algorithm for object identification, we must specify the number of clusters or a stopping condition so that the algorithm stops the clustering and outputs the results after a certain number of clusters have been found.

Determining the number of clusters as a parameter in an object-identification problem is not easy. One reason for this difficulty is that the number of corresponding objects varies considerably from name to name. For example, in the DBLP computer science bibliography[3], which is commonly used as a test collection for object identification, we observed that the number of corresponding full names (clusters) k and the frequency f of abbreviated names obey a power-law distribution: $f(k) = \alpha k^{-\gamma}$ (α and γ are parameters).

In a power-law distribution, a very large number of data items with low values coexist with a few data items with very high values. Thus the average value of the data is meaningless, and there are no "typical" data values. For example, in the data set we used, the average number of full names per abbreviated name is 1.5, but setting the parameter of the number of clusters to 1 (which means doing no clustering) or 2 for all names is not meaningful because that results in very poor performance for names with very many clusters. Therefore, we need to use a different number of clusters for each clustering problem with a distinct ambiguous name.

[1] Kyoto University, Japan, email: oyama@i.kyoto-u.ac.jp
[2] Kyoto University, Japan, email: ktanaka@i.kyoto-u.ac.jp
[3] http://dblp.uni-trier.de/

2 SUPERVISED-LEARNING APPROACH

Previous methods to determine the number of clusters take an "unsupervised" approach and treat each clustering problem independently [1, 2, 3]. In contrast, we take a supervised approach that uses other clustering problems for which we know the true numbers of clusters to predict the number of clusters for an unknown problem. We think this is a reasonable approach for object identification where we solve many similar clustering problems for different names in the same domain. Our approach avoids unnecessary clustering for data sets with one cluster because model-based prediction of the numbers of clusters is used. This is especially effective for object identification when the numbers of clusters follow a power-law distribution and one-cluster problems (problems with no need for clustering) are a large proportion of the problems.

Assume we have pairs of a data set S^j to cluster and the true number of clusters in it, y^j, where the pairs are denoted as $T = \{(S^1, y^1), (S^2, y^2), \ldots, (S^{|T|}, y^{|T|})\}$. Using T as training data, we construct a function f_T that gives a prediction y of the number of clusters for an unknown data set S. We can consider various forms of function f_T. Among them, one of the simplest models is a linear model, $y = \sum_i w_i x_i + b$, where $\{x_i\}$ are explanatory variables that characterize the data set to be clustered, and $\{w_i\}$ and b are parameters determined from the training data T.

The number of clusters should be predicted efficiently. The computational cost of k-means is $O(kn)$ and that of a hierarchical clustering method is $O(n^2)$. Therefore, in practice, the prediction of the number of clusters should be done in linear time with respect to the number of data. Our model should return the number of clusters given a data set cluster, so we need explanatory variables that characterize the statistics of the set of data rather than each datum. In addition, efficiently computing explanatory variables is required. Aggregations of attribute values of the data items to be clustered are good candidates for such explanatory variables. We devised several types of variables that might be correlated with the number of clusters. The explanatory variables we will introduce can be computed in linear time with respect to the number of data items. We can easily compute the value of the aggregated variables by using aggregate functions such as `count()`, `max()`, `min()`, and `avg()`, which are available in most database systems.

We use support vector regression [4] to determine the parameters in the linear model. One difficulty in building a model to predict values from a skewed distribution like a power-law distribution is that there is a large imbalance in the numbers of available training data for different target values. A large portion of the training data is shared by the data items with a target value of 1, and there are relatively few data items with large target values. If we use such training data, there is a risk of obtaining a model that underestimates the target values.

To overcome the problem of imbalance between the numbers of training data for different target values, we introduce a method that successively applies two different models when predicting the number of clusters: (1) One model determines whether a given data set is composed of one cluster or multiple clusters. (2) The other model determines the number of clusters for a data set predicted to be composed of multiple clusters by the previous model. In ecology, a similar two-stage method is used to build a model to predict the abundance of rare species [5], although the learning methods used in each stage are different from ours.

Another extension is that we use a model that is nonlinear to the explanatory variables rather than a linear model. Specifically, we consider a model using combinations of the explanatory variables. Using a higher-order model with large expressive power helps avoid the risk of under-fitting the training data, which sometimes occurs when applying a simple linear model to skewed data. In our implementation, we adopt a kernel trick and use a quadratic polynomial kernel: $k(\boldsymbol{x}, \boldsymbol{z}) = (\langle \boldsymbol{x}, \boldsymbol{z} \rangle + 1)^2$. By using the kernel in support vector learning, we can virtually use the conjunctions of explanatory variables in the model without actually computing the values of conjunctions.

3 EXPERIMENTS

We took the disambiguation of abbreviated author names in a bibliographic database as an example task. From the DBLP data, we randomly selected 2,000 abbreviated names corresponding to more than one paper. We did not use abbreviated names that corresponded to only one paper because there is obviously only one cluster (full name) for them. For each selected abbreviated name, we collected bibliographic data containing the name as an author and computed the value of the following explanatory variables: (1) Number of papers with the target abbreviated author name, (2) Number of different coauthors in the data set, (3) Number of different words appearing in the paper titles, (4) Number of different journals or conference proceedings in which the papers are published, (5) Difference between publication years of the newest and oldest papers, (6) Standard deviation of publication years of papers in the data set, (7) Frequency of last names used in abbreviated names in the database, (8) Percentage of abbreviated names with a particular letter among the abbreviated names.

We applied 10-fold cross validation. We used SVM^{light}[4], which implements support vector regression to build the regression models as well as binary support vector machines used in building two-stage models. As the metric, we used the root mean square error (RMSE) between the true number of clusters (full names) and the predicted number of clusters given by a model.

We compared the Caliński and Harabasz (C&H) method [1], the Hartigan method [2], a method using an average threshold, x-means [3], the basic learning-based method (Linear (1 stage)), a two-stage method (Linear (2 stages)), nonlinear regression using a polynomial kernel (Polynomial (1 stage)), and a two-stage method using a polynomial kernel in each stage (Polynomial (2 stages)). For C&H, Hartigan, and x-means, we simply applied the methods for the clustering problems in the test sets and did not use the training sets. For the method using an average threshold, we applied the single-linkage method for each clustering problem in the training set and calculated the average of the thresholds that resulted in the true numbers of clusters. We then applied the single-linkage method to each clustering problem in the test sets and determined the number of clusters by using the average threshold as the clustering-stopping condition.

The overall RMSE for each method is shown in Table 1. The four learning-based methods outperformed the other methods. Among the four learning-based methods, the two-stage model and the model with the polynomial kernel outperformed the basic model, and their combination gave the results with the smallest errors.

Table 1. RMSE for each method

C & H	3.063	Linear (1 stage)	1.819
Hartigan	2.279	Linear (2 stages)	1.490
Threshold	2.231	Polynomial (1 stage)	1.145
X-means	2.585	Polynomial (2 stages)	1.114

4 CONCLUSION

We described a supervised, model-based approach to predicting the number of clusters in a data set, which is more efficient and accurate than existing approaches. In addition, it enables us to avoid unnecessary clustering for one-cluster problems, which are a large proportion of the problems. As explanatory variables used in the prediction model, we used aggregated attribute values of the data set to be clustered, which can be computed efficiently. We described a basic learning-based method using a linear model as well as two extended methods: a two-stage method and a method using combinations of explanatory variables. Experimental results in author disambiguation showed that our learning-based methods outperformed existing methods and that the two extensions improved the performance of the basic linear model.

ACKNOWLEDGMENTS

This work was supported in part by Grants-in-Aid for Scientific Research (Nos. 18049041 and 19700091) from MEXT of Japan, a MEXT project entitled "Software Technologies for Search and Integration across Heterogeneous-Media Archives," a Kyoto University GCOE Program entitled "Informatics Education and Research for Knowledge-Circulating Society," and a Microsoft IJARC CORE4 project entitled "Toward Spatio-Temporal Object Search from the Web."

REFERENCES

[1] T. Caliński and J. Harabasz, 'A dendrite method for cluster analysis', *Communications in Statistics*, **3**(1), 1–27, (1974).
[2] J. A. Hartigan, *Clustering Algorithms*, Wiley, 1975.
[3] D. Pelleg and A. Moore, 'X-means: Extending K-means with efficient estimation of the number of clusters', in *Proceedings of ICML 2000*, pp. 727–734, (2000).
[4] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.
[5] A. H. Welsh, R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer, 'Modelling the abundance of rare species: Statistical models for counts with extra zeros', *Ecological Modelling*, **88**, 297–308, (1996).

[4] http://svmlight.joachims.org/