Distributed Data Clustering

Abdelhamid Bouchachia

Universität Klagenfurt, Institut für Informatik-Systeme
Universitätsstrasse 65, A-9020 Klagenfurt, Austria
hamid@isys.uni-klu.ac.at

Abstract. To make effective use of distributed information, it is desirable to allow coordination and collaboration among various information sources. This paper deals with clustering data emanating from different sites. The process of clustering consists of three steps: find the (local) clusters of data at each site; find (higher) clusters from the union of the distributed data sets at the central site; and finally compute the associations between the two sets of clusters. The approach aims at discovering the hidden structure of a multi-source data and assigning unseen data points coming from a site to the right higher cluster without any need to access their feature values. The proposed approach is evaluated experimentally.

1 Introduction

Due to advances in communications technology, the number of distributed information sources accessible to a seeker has grown rapidly. To make effective use of this information, it is desirable to allow coordination and collaboration among various sources. This is especially relevant in public administration for instance. But, because the society is becoming more and more dependent on information, new constraints such as individual privacy and corporate confidentiality arise. Such notions and others such as efficiency have been thoroughly discussed in the knowledge discovery literature [2][5][7][8]. Clustering offers an opportunity to deal with various constraints in an appropriate way. It allows categorizing objects (e.g customers) and understanding the correlation of sources (e.g. services provided by institutions) where objects emanate from. In this context, the term correlation stands for collaboration. Each data set emanating from a site provides just a piece of information about the customer and the task of a central site (where data is gathered) is then to take advantage from the whole distributed knowledge. However, the central site must take care of the privacy of data coming from a given site. An institution's data must not be disclosed to other institutions. The present work deals with the problem of clustering distributed data emanating from different sites. It aims at finding the contribution of each individual data set in building clusters from the union of the data sets taken as a whole. To take into account the constraints mentioned earlier, an intermediate solution is suggested. In fact, the idea is to communicate a sample of data only after agreement upon the privacy between the central site and the distributed sites. Once gathered from different sites, the data have to be formatted to come up with a single data set. Thus issues related to the structure and the data points have to be taken care of at the central site [3]. Here, the data sets are simply merged together so that the structure of the resulting data set consists of features coming from individual data sets. Actually we need the data from distributed sites just for a preliminary phase, that to compute the contribution of each individual data set in defining the output which is a set of higher clusters at the level of the central site.

2 Clustering approach

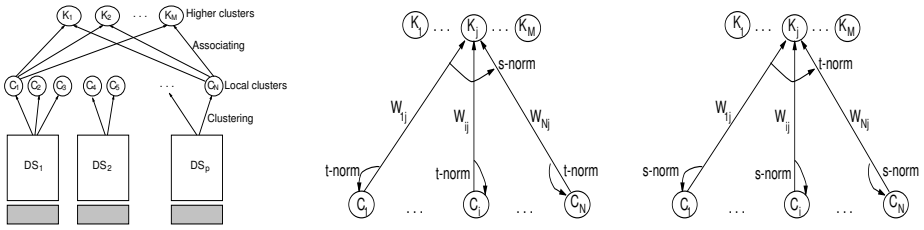
To compute the level of contribution of each data set in building the clusters at the central site on one hand and to be able to categorize new data points coming from distributed data without having access to the values of their features on the other hand, we proceed in three steps as follows: **(a)** the first step consists of building clusters C_i (called local clusters) in each data set, **(b)** Then, clusters K_j (called higher or global clusters) are built from the data set resulting from the union of individual data sets at the central site (In this work, the standard fuzzy C-Means (FCM) [1] algorithm is applied for clustering data) and **(c)** after generating clusters at both levels, the aim is then to discover the type of relationship between local and higher clusters. Figure (1a) visualizes the three steps. While the first two steps are easily performed, the last one needs more investigation and development. Being an issue of mapping, the association between the local and higher clusters is modelled using a learning mechanism. The idea is not only to compute associations between local and higher clusters (i.e., the contribution of each individual data set in the clustering results at the higher level) but also for assigning unseen data points in the future without the need to access their feature values. Thus two phases are required: a *training phase* and an *operational phase*. The first phase allows finding the associations. In the second phase, clustering unseen data points using the associations computed is performed. The learning process is based on the gradient descent algorithm whose details are discussed below. Based on the idea provided in [6] the relationship between clusters can occur in two forms; namely *abstraction* and *specialization*. If the number of clusters in the higher level is smaller than that of the lower level, we have an abstraction (or aggregation) case, otherwise it is a specialization case. In a different context, Pedrycz [6] used fuzzy Boolean operators to handle the relationship between clusters. A cluster is simply a fuzzy granule (set) whose elements are the data points represented by their membership grades. This idea will be adopted here, but with further developments. In the case of abstraction higher, clusters are realized as a union of local clusters, while in the case of specialization, a higher cluster is the conjunction (overlap) of local clusters. Note that the membership grades of data points to clusters come in the form of the partition matrix. Hence, it is easy to apply fuzzy operations (union "or" and conjunction "and") to compute the membership grade of data points to the higher clusters. In the sequel, each of these two options is discussed.

2.1 Abstraction and Specialization

Higher clusters are constructed by means of a union of local clusters. However, simple union cannot reflect the whole association because a higher cluster might be the result of the union of local clusters with some additional information denoted by W. Therefore, a higher cluster is represented as:

$$K_j = (C_1 \wedge w_1) \vee (C_2 \wedge w_2) \vee \cdots \vee (C_N \wedge w_N) \tag{1}$$

where the additional information is expressed by means of another fuzzy operation which is the fuzzy "and". Of course, the fuzzy "or" and "and" can be generalized to fuzzy t- and s-norms(see Fig. 1b). Now, clusters look as nodes and associations as connections with weights, we might be interested to look at the problem of finding those



(a) Association local-higher clusters (b) Simple abstraction (c) Simple specialization

Fig. 1. Relationship between Local and global clusters

connection weights. Hence, a connectionist approach can be applied. In fact, we have a special neural network that consists of two layers where the output nodes are an OR-like nodes. The basis function is therefore the fuzzy OR. An output node is expressed as:

$$y = \underset{i=1}{\overset{N}{S}} w_i T x_i$$

where, x_i is the input, w_i is the connection weight, S is the s-norm, and T is the t-norm. Having this architecture, we can develop a learning algorithm that finds the connection weights W. If we consider that the input is the set of vectors corresponding to the membership degrees of the patterns to local clusters, the output will be the set of vectors corresponding to the membership degrees of the patterns to the higher clusters. To find the weights the gradient descent method is applied. If we consider probabilistic norms [4], and develop the gradient descent method to find the weights, we get the learning rule:

$$w_j^{(t+1)} = w_j^{(t)} + 2\eta x_j^p (t_j^p - y_j^p)(1 - R_j) \tag{2}$$

where: $$R_j = \underset{i \neq j}{\overset{N}{S}} w_i T x_i^p,\ T(a,b) = ab,\ S(a,b) = a + b - ab$$

On the other hand, specialization is the operation by which two clusters are joined by an 'and' operator. The operation of 'and'ing clusters leads to more specific clusters with shared knowledge. The association between local and higher clusters can be viewed as in Figure 1c. Following the same steps as in the abstraction operation, the learning algorithm will rely on the learning rule:

$$w_j^{(t+1)} = w_j^{(t)} + 2\eta Q_j (t_j^p - y_j^p)(1 - x_j^p) \tag{3}$$

where that: $$Q_j = \underset{i \neq j}{\overset{N}{T}} w_i S x_i^p$$

3 Evaluation

To evaluate the approach a real-world data set about breast cancer[1] is used. It consists of 8 features. The data set DS which consists of 8 features will be divided into 2 data

[1] Detailed description can be found in http://www.ics.uci.edu/ mlearn/MLRepository.html

sets $DS_1(f_1,f_2,f_3,f_4)$, and $DS_2(f_5,f_6,f_7,f_8)$. For the evaluation purpose, we assume that each data set is coming from a site. A set of 300 data points of DS are used to train the associator. The learning parameters, namely the number of iterations and learning rate η (see Eq.2), are assigned the values 2000, and 0.005 respectively from results of a preliminary experiment. The associator was evaluted using a testing sample of data points (grey part in Figure 1a). We used a sample of 150 data points of the breast data to test the efficiency of the associator for both cases of abstraction and specialization. Concerning abstraction, the number of points correctly assigned is 146, hence a success rate of 97.33%. For the specialization case, the number of association successes is 141, i.e., a success rate of 94%. It is noticeable that the approach provides a high assignment accuracy for both abstraction and specialization cases. In fact, it is able to assign unseen data points to the right higher cluster. The central site gets just the membership degrees of new data points and will be able in the future to assign them to clusters using the associator. This means that the distributed sites communicate to the central site only the partition matrix resulting from the clustering process performed locally. Raw data will not be needed any more. Hence, the data confidentiality is preserved.

4 Conclusion

The approach presented here suggests the use of clustering for distributed data to deal with the constraints of confidentiality and privacy. An associator is computed allowing to infer the contribution strength of individual data sets in bearing the semantic content of all data gathered at the central site. Further investigations are required to generalize the finding. For instance, the data used here is a heterogeneous one, i.e., individual data sets have different structure (features) and hence their merger is straightforward. It sounds very interesting to make the approach applicable to (partially or fully) homogeneous data where a subset of features are common to some (or to all) individual data sets and this will be assessed in the future.

References

1. J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.
2. H. Kargupta and et al. editors. *Proc. Workshop on Distributed Data Mining. In Conj. with the* 4^{th} *Inter. Conf. on Knowledge Discovery and Data Mining*. New York, 1998.
3. H. Kargupta, B. Park, D. Hershberger, and E. Johnson. *Advances in Distributed and Parallel Knowledge Discovery*, chapter Collective Data Mining: A New Perspective Toward Distributed Data Mining. MIT/AAAI Press, 1999.
4. C. Lin and C. Lee. *Neural Fuzzy Systems*. Prentice Hall, 1996.
5. R. Páircér, S. McClean, and B. Scotney. Automated Discovery of Rules and Exceptions from Distributed Databases Using Aggregates. *Proc. of the* 3^{rd} *European Conf. on Principals of Data Mining and Knowledge Discovery*, pages 156–164, 1999.
6. W. Pedrycz and G. Vukovich. Abstraction and Specialization of Information Granules. To appear in IEEE Trans. on Systems Man and Cybernetics.
7. S. Sarawagi and S.H. Nagaralu. Data Mining Models as Services on the Internet. *SIGKDD*, 2(1):24–28, 2000.
8. S. Stolfo, A. Prodromidis, S. Tselepis, W. Lee, D. Fan, and P. Chan. JAM: Java Agents for Meta-Learning over Distributed Databases. *Proc. of the* 3^{rd} *Inter. Conf. on Data Mining and Knowledge Discovery*, pages 74–81, 1997.