OBSERVATION ON CLUSTERING MULTI REPRESENTED OBJECTS TO CALCULATE QUALITY

M. Parvathi, MCA,M.Phil., Ph.D

Director Science, Head, Dept of Computer Applications, Senthamarai College of Arts and Science, Madurai-625 016.

ABSTRACT: *Recent technological advances have tremendously increased the amount of collected data. Besides the total amount of collected information, the complexity of data objects increases as well. With the growing amount of data the complexity of data objects increases. Modern methods of KDD should therefore examine more complex objects than simple feature vectors to solve real-world KDD applications adequately. To analyze these data collections, new data mining methods are needed that are capable to draw maximum advantage out of the richer object representations. This paper contributes to the field of data mining of complex objects by introducing methods for clustering and classification of compound objects. The area of KDD deals with analyzing large data collections to extract interesting, potentially useful, so far unknown and statistical correct patterns. The data objects to be analyzed are of complex nature. Thus, they are not represented in the best possible way by the common approach using feature vectors. So, data mining algorithms have to handle more complex input representations. KDD is necessary to analyze the steady growing amount of data caused by the enhanced performance of modern computer systems.*

Multi-represented objects are constructed as a tuple of feature representations where each feature representation belongs to a different feature space. The ***goal*** *of clustering multi-represented objects is to find a meaningful global clustering for data objects that might have representations in multiple data spaces. This paper contributes to the development of clustering and classification algorithm that employ more complex input representations to achieve enhanced results. Therefore, the paper introduces solutions for real-world applications that are based on multi-represented object representations. To analyze multi-represented objects, a clustering method for multi-represented objects is introduced that is based on the projected based clustering algorithm. This method uses all representations that are provided to find a global clustering of the given data objects. To map new objects into ontology a new method for the hierarchical classification of multi-represented objects is described. The system employs the hierarchical structure of the efficient classification method using support vector machines.*

KEYWORDS: **Complex objects, Entropy, KDD, Support Vector Machine, Threshold**

1. INTRODUCTION

A multi-represented object consists of a tuple of feature representations. Each feature representation belongs to a different feature space and represents another view on the object. Many important areas of KDD are concerned with finding useful patterns in large collections of complex objects. Images, biomolecules or CAD parts are only some examples of complex objects that are in the center of interest of many researchers. However, the more complex a type of object is the more feature transformations exist that try to extract relevant features and construct a meaningful object representation. All of these feature transformations are well suited for different applications and treat a data object from another point of view.

For data mining, the existence of multiple feature transformations is often problematic because it is not clear which of the representations contains the features that are needed to achieve the desired results. Thus, the selection of a feature transformation is often a difficult and crucial decision that strongly influences the resulting patterns. Clearly, incorporating all available feature transformations offers a more complete view of a data object and minimizes the hazard that the information that is necessary to derive meaningful patterns are not contained in the object representation. On the other hand, considering too many aspects of an object is often problematic as well. The found patterns are often very complicated and lose generality. Furthermore, the efficiency of the data mining algorithms suffers strongly since much more features have to be processed.

2. PREVIOUS STUDY

Most of the algorithms are designed for one feature space and one distance function to represent the data objects. Thus, to apply these algorithms to multi-represented data, it is necessary to unite the representations into one common feature space.

A similar setting to the clustering of multi-represented objects is the clustering of heterogeneous or multi-typed objects in web mining. In this setting, there are also multiple databases, each yielding objects in a separated data space. Each object within these data spaces may be related to an arbitrary amount of data objects within the other data spaces. The framework of reinforcement clustering employs an iterative process based on an arbitrary clustering algorithm. It clusters one dedicated data space while employing the other data spaces for additional information. It is also applicable for multi-represented objects. The best of our knowledge reinforcement clustering is the only other clustering algorithm directly applicable to multi-represented objects.

The setting of reinforcement clustering is to cluster the data within one data space while using the related data spaces for additional information. Since the results may vary for different starting representations, the application of reinforcement clustering is problematic. It is unclear how much iteration are needed until a common clustering for all representations is found and if the algorithm reaches a common clustering at all for an arbitrary number of iterations.

3. CLUSTERING MULTI-REPRESENTED OBJECTS

A multi-represented object is described by a set of representations where each representation belongs to a different data space.

3.1. WHY MULTI-REPRESENTED OBJECTS?

- The existence of several useful feature transformations that model different, important aspects of the same data objects, e.g. shape and color of an image.
- The occurrence of multi-represented objects is the existence of different measuring techniques for

an object.

- If the databases are integrated into a global data collection, the global view contains representations from each of the source database.

Clustering methods are targeted at finding groups of similar objects in one type of object representation using one distance function. In this paper, projected-based clustering of multi-represented objects is examined.

To cluster multi-represented data, using our new clustering methods would require to restrict the analysis to a single representation or to construct a feature space comprising all representations. However, the restriction to a single feature space would not consider all available information and the construction of a combined feature space demands great care when constructing a combined distance function. Since the distance functions best-suited for each representation might not even provide the same value set, it is difficult to find a proper combination that gives a meaningful distance. The main problem here is several data objects might not provide all possible representations. If the representation is expensive and time consuming, we have to use three dimensional models. In this case, the combined distance function would need to handle missing representations adequately. Since many clustering algorithms are based on similarity queries, the use of index structures is usually very beneficial to increase the efficiency, especially for large data sets.

3.2. Union of different representations

This variant is especially useful for sparse data. In this setting, the clustering in each single representation will provide several small clusters and a large amount of noise. Simply enlarging would relief the problem, but the separation of the clusters would suffer. The union-method assigns objects to the same cluster if they are similar in at least one of the representations. Thus, it keeps up the separation of local clusters, but still overcomes the sparsity. For sparse data the union method that is based on the assumption that an object should be a core object, if k objects are found within the union of its localε -neighborhoods.

3.3. Intersection of different Representations

The intersection method is well suited for data containing unreliable representations. In those cases, the intersection-method requires that a cluster should contain only objects which are similar according to all representations. Thus, this method is useful if all different representations exist but the derived distances do not adequately mirror the intuitive notion of similarity. This method is used to increase the cluster quality by finding purer clusters. This method was introduced for data where each local representation yields rather big and unspecific clusters. This method requires that at least k objects are within the intersection of all local ε-neighborhoods of a core object. Thus, this method is much more restrictive.

4. PROPOSED CLUSTERING ALGORITHM

At the beginning of the clustering process, each object forms a singleton cluster. The dimensionality and relevance thresholds din and Rmin are initialized to their tightest values. For each cluster, the dimensions that satisfy the threshold requirements are selected. The right score between each pair of clusters is then calculated. Only the merges that form a resulting cluster with dmin or more selected dimensions are qualified. The other merges are being ignored.

The new algorithm was developed to find out the union of different objects. If the objects are in k number then it is to be clustered together based on their distance that is minimum radiant distance will be calculated and clustered. Check the remaining objects and continue the same testing and clustering The algorithm repeatedly performs the best merge according to the MS scores of the qualified merges. In order to efficiently determine the next best merge, right scores are stored in a cache. After each merge, the scores related to the merged clusters are removed from the cache, and the best scores of the qualified merges that involve the new cluster are inserted back. The selected dimensions of the new cluster are determined by its members according to Rmin. If a dimension is originally not selected by both merging clusters, it must not be selected by the new cluster. However, if a dimension is originally selected by one or both of the merging clusters, it may or may not be selected by the new cluster

Procedure SelectDim

```
For each dimension vj
If RIj≥ Rmin and ValidRel(CI , vj)
Select vj for CI
End f the
```

Procedure SelectDimNew

```
For each dimension vj {
If $R^{*}$ Ij≥ Rmin and ValidRel (CI1 , vj) and ValidRel (CI2 , vj)
Select vj for CI3
} End
```

Procedure ValidRel

```
lowv = max(xIj - 2$\delta$Ij , minIj)
highv = min(xIj + 2$\delta$Ij, maxIj)
If mean frequency of the bins covering [lowv, highv] < mean frequency of all bins
return FALSE
Else
return TRUE
End
```

Procedure UpdateScoreCache

```
Delete all entries involving CI1 and CI2 from cache
Foreach cluster CI4 ≠ CI3 do {
CI5 = CI3 Ụ CI4
SelectDimNew(CI5 , Rmin)
If dI5 ≥ dmin
Insert MS(CI3 , CI4) into score cache
} End
```

Whenever the cache becomes empty, there are no more qualified merges at the current threshold level. The thresholds will be loosened linearly according to the formula in lines 2 and 3 of Thabclus algorithm. Further rounds of merging and threshold loosening will be carried out until a target number of clusters remains, or the thresholds reach their baseline values and no more qualified merges exist.

The MS score between each clustered object and each cluster is computed based on the final threshold values when the hierarchical part ends. After computing all the scores, each of the object is assigned to the cluster with the highest MS score. The process repeats until convergence or a maximum number of iterations reached. When two clusters merge to form a new cluster, two deletions and one insertion are required. The whole algorithm requires no user parameters in guiding dimension selection or right score calculation, so it can easily be used in real applications. The high usability is attributed to the dynamic threshold loosening mechanism, which relies on the hierarchical nature of Thabclus. Thabclus is especially suitable for applications where accuracy is the first priority and the datasets are of moderate sizes.

To extend the algorithm, we redefined the core object property by the union and the intersection method. The idea of the intersection method is that in order to be a core object a data object should be placed in a dense region in all representations. Thus, it is well suited for applications in which the proximity of two object representations is necessary but not sufficient to indicate the proximity of the original objects. The union method is enough that an object is placed in a dense region with respect to all representations.

Comparability of local results and data objects: To combine the representations for deriving a global pattern, the meaning of each representation has to be made comparable. To achieve comparability there exist several approaches.

Joined Data Spaces: By simply joining the data spaces into a single high dimensional vector space, we lose the information that different features are derived from different representations.

Combined distance functions:

Here to use the local specialized distance functions in each representation and combine the local distance values to a global distance.

Recombination of Local Patterns:

Here the local patterns are recombined to get a global result. Since the distance between two objects can be considered as deriving a local pattern. The advantage is that local patterns provide a higher level of abstraction and are rather independent from the data distribution in the single representations.

Semantics of the Representations

Here the problem is the meaning of a representation. Some representation might contain less reliable information than others. In many applications, the meaning of each representation might be unknown or difficult to describe in advance. In these cases, finding the correct semantics should be estimated by the data mining algorithm. Another problem is the relationship of the given application to each of the representations. Since the combination can be optimized with respect to the correct classification of the training objects, the semantics can be discovered automatically.

In our experimental evaluation, we introduced an **entropy based quality measure** that compares a given clustering with noise to a reference clustering. Employing this quality measure, we demonstrated that the union method was most suitable to overcome the sparsity of a given data set. To demonstrate the ability of the intersection method to increase the cluster quality, we applied it to a set of images using two different models.

5. PERFORMANCE EVALUATION

The noise ratio increased for the intersection-method, because here very similar images are clustered together. When clustering each single representation, a lot of additional images were added to the corresponding cluster. The experimental evaluation demonstrated that our new method is capable to derive more meaningful clustering compared to other clustering methods.

Deriving Meaningful Groupings

The first set of experiments was performed on protein data that is represented by amino acid sequences and text descriptions for PAX6 Gene. So we employed entries of the SWISS-PROT protein database and transformed each protein into a pair of feature vectors. To compare the derived feature vectors, we employed the Euclidian distance. To process text documents, we rely on projecting the documents into the feature space of relevant terms. We chose 100 words of medium frequency as relevant terms and employed distance.

We chose the entropy as measure for cluster quality because our reference clustering does not provide real clusters but classes. The entropy considers a cluster as good as long as its objects belong to the same class. The effect that the ideal cluster consists of a clustering where each object corresponds to its own cluster, is avoided. A clustering providing an extraordinary amount of noise can contribute only the percentage of clustered objects to the quality.

To relate the quality of the clustering achieved by the new method, we compared it to four alternative approaches. First, we clustered text and sequences separately, using only one of the representations. The second approach combines the features of both representations into a common feature space and employs the distance to relate the resulting feature vectors.

The improvement rates of the cluster quality for the union method were between 38% and 65%. Furthermore, the noise ratio for each data set was between 14% and 35%, indicating that the main portion of the data objects belongs to some cluster. The intersection method performed comparably well, but was too restrictive to overcome the sparseness of the data as good as the union-method. Employing this quality measure, we demonstrated that the union method was most suitable to overcome the sparsity of a given protein data set.

6. CONCLUSIONS

In projected based clustering of multi-represented objects, there exist other directions of clustering that are suitable for other kinds of applications. To exploit multiple representations in these applications, the use of multi-represented objects in combination with other directions of clustering yields many interesting aspects. When classifying multi-represented objects into large class hierarchies, the use of SVMs tends to be very inefficient. We plan to develop new classifiers for multi-represented objects that use kNN classification and are well suited for very large class spaces. Then, we plan to combine the introduced methods for data mining in multi-represented into a general approach for data mining in compound objects. For this approach, an object could be constructed arbitrarily of concatenations and sets of other feature representations like graphs, tree and feature vectors. Especially in the area of protein databases many representations in a multi-represented view might be modelled more precisely e.g. the 3-D structure. Thus, the use of even richer protein descriptions could yield an even better approach to clustering and classification of this kind of data.

REFERENCES

[1] H.-P. Kriegel, P. Kr¨oger, A. Pryakhin, and M. Schubert. ”UsingSupport Vector Machines for Classifying Large Sets of Multi-Represented Objects”. In Proc. SIAM Int. Conf. on Data Mining, Lake Buena Vista, FL, USA, pages 102–113,2014.

[2] K. Kailing, H.-P. Kriegel, A. Pryakhin, andM. Schubert. ”ClusteringMulti -Represented Objects with Noise”. In Proc. 8thPacific-Asia Conf. on Knowledge Discovery and Data Mining, Sydney, Australia, pages 394–403, 2013.

[3] T. Joachims. ”Text Categorization with Support Vector Machines: Learning with Many Relevant Features”. In Proc.10th European Conf. on Machine Learning, Chemnitz, Germany, Lecture Notes in Computer Science (LNCS), Springer, pages 1398: 137–142, 2013.

[4] T. Zhang, R. Ramakrishnan, and M. Livny. ”BIRCH: An Efficient Data Clustering Method for Very Large Databases”. InProc. ACM SIGMOD Int. Conf. on Management of Data, Montreal, Canada, pages 103–114, 2012.

[5] Y. Yang and P.O. Pederson. ”A comparative study on feature selection in text categorization”. In Proc. 14th Int. Conf. on Machine Learning, Nashville, TN, USA, pages 412–420, 2012.

[6] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W. Ma. ”Re-CoM: reinforcement clustering of multi-type interrelated data objects”. In Proc. 26th ACM SIGIR Conf. on Research and Development in Information Retrieval, Toronto, Canada, pages 274–281, 2013.

[7] J.T.L. Wang, Q. Ma, D. Shasha, and C.H. Wu. ”New techniques for extracting features from protein sequences”. IBM Systems Journal, 40(2), 2011.

[8] S. Vaithyanathan, J. Mao, and B. Dom. ”Hierarchical Bayesfor Text Classification”. In Proc. Int. Workshop on Text and Web Mining, Melbourne, Australia, pages 36–43, 2012.