Clustering Zebrafish Genes Based on Frequent-Itemsets and Frequency Levels

Daya C. Wimalasuriya,[1] Sridhar Ramachandran,[2] and Dejing Dou[1]

[1] Department of Computer and Information Science
University of Oregon, USA
{dayacw,dou}@cs.uoregon.edu
[2] Zebrafish Information Network
University of Oregon, USA
sramacha@uoregon.edu

Abstract. This paper presents a new clustering technique which is extended from the technique of clustering based on frequent-itemsets. Clustering based on frequent-itemsets has been used only in the domain of text documents and it does not consider frequency levels, which are the different levels of frequency of items in a data set. Our approach considers frequency levels together with frequent-itemsets. This new technique was applied in the domain of bio-informatics, specifically to obtain clusters of genes of zebrafish (Danio rerio) based on Expressed Sequence Tags (EST) that make up the genes. Since a particular EST is typically associated with only one gene, ESTs were first classified in to a set of classes based on their features. Then these EST classes were used in clustering genes. Further, an attempt was made to verify the quality of the clusters using gene ontology data. This paper presents the results of this application of clustering based on frequent-itemsets and frequency levels and discusses other domains in which it has potential uses.

1 Introduction

1.1 Clustering Based on Frequent-Itemsets

Clustering based on frequent-itemsets is recognized as a distinct technique and is often categorized under frequent-pattern based clustering methods [10]. There are many clustering techniques categorized under this theme and some of them are not based on frequent-itemsets; they are generally based on the frequent patterns observed in some of the dimensions in high-dimensional data. For instance, the pClustering method [12] which performs clustering by pattern similarity in microarray data analysis is generally identified as a frequent-pattern based clustering technique. While all these clustering techniques clearly have something in common in terms of discovering clusters based on frequent patterns, the patterns involved are quite different in different clustering techniques. In techniques such as pClustering these patterns refer to the patterns observed in the values of some dimensions for a set of objects whereas in techniques based on frequent-itemsets the patterns are the frequent-itemsets.

Z.-H. Zhou, H. Li, and Q. Yang (Eds.): PAKDD 2007, LNAI 4426, pp. 912–920, 2007.

Clustering techniques based on frequent-itemsets have been hitherto applied almost exclusively in the domain of text documents. In a pioneering work in this area presented by Beil *et al.* [9], two algorithms for discovering clusters of documents based on frequent terms (a sequence of characters separated from other terms by delimiters) they contain have been developed. The two algorithms presented in this paper are based on the concept of assigning the objects that have a frequent-itemset to a cluster. One of the main advantages of this method is that it provides an inherent definition for the clusters, in terms of the frequent-itemsets they have. Such a meaning is generally not provided by other clustering techniques. However, the clusters defined in this manner can be overlapping and non-exhaustive (not covering all the objects) and the algorithms mentioned above are carefully designed to overcome these issues.

The clustering technique used in this work presents an improvement over the basic frequent-itemset based clustering technique because it takes in to consideration the frequency levels of items in a record or an object. Typically, the relative frequency level of an item within a record or an object is not considered in identifying frequent-itemsets; frequent-itemsets are identified based on the items that are observed in a number of objects or records higher than a predefined threshold. For instance, in a data set consisting of 1000 purchase records and with the use of a cut-off limit of 10% , frequent-itemsets are identified as the items that appear in at least 100 records. While this is sufficient in most situations, the frequency of items within a record is also important in some situations. For example, if each purchase record contains at least 100 units, the quantity or the number of times an item appears in each record is also important; in addition to identifying the itemset A,B,C as being frequent, identifying that 30 units of A, 10 units of B and 5 units of C are frequent provides more information. When using these frequent-itemsets for clustering, paying attention to the relative frequencies of items within records in this manner provides more insight in to common characteristics of the objects or records of the cluster. This served as a basis in developing the clustering technique used here.

1.2 Genes and Expressed Sequence Tags (EST)

A gene can be uniquely described by its sequence of nucleotides, which can be thousands of nucleotides long. There are several techniques that are used to identify genes and the use of Expressed Sequence Tags (ESTs) is one such technique. ESTs provide a snapshot of the DNA that is expressed in a given tissue of a eukaryotic organism at a given time. This is in accordance with the Central Dogma of Molecular Biology, which states that DNA first has to be converted in to RNA through the process of transcription and then the RNA has to be converted to proteins, which do the actual work of altering a cell's chemistry, through the process of translation. ESTs are typically short and restricted to about 300 - 500 nucleotides. Since they indicate the regions of DNA that have been transcribed to RNA, they can be used to identify genetic material that are active in a particular situation. Further, contiguous blocks of DNA can be assembled using ESTs and these blocks can be used to identify genes.

Several organizations keep records of the genes and ESTs discovered by researchers and make them publicly available through online databases. One such organization is GenBank [2], which provides access to different types of genetic data. UniGene [6], which is a part of GenBank, provides details on the genes that have been identified based on ESTs and other techniques. Further, it is possible to directly download the data that show how genes have been constructed using ESTs, for each species as a single file. One such file is available for zebrafish. In essence, it lists out the ESTs of each gene of zebrafish. The details of ESTs can be obtained from GenBank, which provides a list of files that contain the details of all the ESTs found in different species. Each record of a file provides comprehensive details on an EST including its nucleotide sequence and the tissue type and/or development stage it was observed.

The main objective of the research was to identify groups of genes based on the similarity of their EST makeup. Since the ESTs provide an idea of the active genetic material in a particular tissue type at a particular development stage of the organism, it is reasonable to assume that genes that have a highly similar EST make up are active in the same tissue type and/or the same development stage. This would probably indicate that such genes may have similar functions in particular tissues or are involved in a common biological process. There have been previous work on analyzing the EST makeup of genes. GEPIS [4] integrates EST and tissue source information to compute gene expression patterns in normal and tumor samples. EST miner [5] is another work in this area.

Any given EST is a part of one and only one gene since genes are identified based on a contiguous sequence of nucleotides. An EST constitutes a section of the nucleotide sequence of the DNA of an organism. As such the unique identification numbers of ESTs can not be used in clustering genes since a given identification number is found in only one gene. Hence the approach adopted was to classify the ESTs in to a group of classes according to the tissue type and the development stage they are found and then use these EST classes in clustering genes.

2 Related Work

Although the concept of discovering clusters based on frequent-itemsets is generic, it has clearly been associated with clustering of text documents. Beil *et al.* [9] use the term “term set” to highlight the fact that the items in text documents are terms and defines the technique as “frequent term-based clustering” instead of “frequent-itemset based clustering”. This name has been used by others as well [10]. According to this technique, when a frequent-itemset is identified the set of records or objects that have the frequent-itemset in concern becomes a potential cluster. Different variations of this clustering technique such as Frequent-Term based Clustering (FTC) and Hierarchical Frequent Term-Based Clustering (HFTC) described in [9] identify the final set of clusters from these potential clusters in different ways. In its pure form, the potential clusters identified using frequent-itemsets are overlapping and non-exhaustive.

Some techniques such as FTC ensure that final clusters are not overlapping and that they cover all the objects (records) making them more consistent with the functionality of classical clustering techniques.

Our work demonstrates the use of a clustering technique based on this method in a very different domain. This shows that this clustering technique is a more generic technique. Further since our technique takes frequency levels into consideration in addition to frequent-itemsets, it might be capable of discovering better clusters. Previous work has shown that paying more attention to the structure of documents would result in better clustering techniques. For instance, Li and Chung [11] have implemented a text document clustering technique, which shows a better performance, by considering frequent word sequences in documents. In addition, this work can be related to other attempts to discover clusters using some kind of summarization of frequent-patterns, such as the research in [13].

3 Our Approach

To identify groups of genes based on the similarity of their EST makeup, we designed a new clustering algorithm based on frequent-itemsets and frequency levels. It is presented below:

Algorithm 1. Clustering based on frequent-itemsets and frequency levels

```
1: Determine a cut-off frequency level for the data set.
2: Identify items to be used in frequent-itemset mining from the records of the data
   set using the cut-off frequency level.
3: Determine a threshold to be used in frequent-itemset mining.
4: Identify frequent-itemsets using the apriori algorithm or the FP-growth technique.

5: Extract meaningful frequent-itemsets.
6: Identify the records that have the meaningful frequent-itemsets by scanning the
   data set for each meaningful frequent-itemset.
7: Present the records that share a meaningful frequent-itemset as a cluster and define
   the meaning of the cluster in terms of the frequent-itemset.
```

In steps 1 and 2 the concept of frequency levels is used to identify items in records, to be used in frequent-itemset mining later. This requires a cut-off frequency level as a parameter. For a particular item of an record, the number of items to be used in frequent-itemset mining is determined by performing integer division (or division followed by floor operation) on its frequency within the record using the cut-off frequency level. This can be expressed as follows.

Let the total count of items in the record be n.
Let count of item i in the record be m. ($m \leq n$)
Let cut-off frequency be c. ($c < 1$)
$\therefore$ number of items of i to be used in frequent-itemset mining $= \lfloor (m/n)/c \rfloor$

A numbering scheme together with a special character is used to represent the fact that the items identified in this manner relate to the same item in the actual record. Assuming that the special character is `$` the following would represent 4 items to be used in frequent-itemset mining, all based on item `A`.

```
A$1, A$2, A$3, A$4
```

In steps 3 and 4, a standard frequent-itemset mining technique is employed on the items identified in the previous steps. This is based on discovering longer frequent-itemsets using shorter frequent-itemsets. Both apriori algorithm and FP-growth technique can be used here. The same method is used in Hierarchical Frequent Term-Based Clustering (HFTC) technique described in [9].

Step 5 extracts the meaningful frequent-itemsets from the set of all frequent-itemsets discovered in step 4. This is necessary because the number included in an item according to the numbering scheme discussed above has a meaning. For example, assuming that the cut-off frequency is 10%, `A$2` represents the second 10% step within the frequency of `A` in the record in concern. This makes some frequent-itemsets meaningless. For instance, if a group of genes have 30% of ESTs of class `A`, all the following are identified as frequent-itemsets, assuming that a cut-off percentage of 10% is used.

```
{A$1},{A$2},{A$3},{A$1, A$2},{A$2, A$3},{A$1, A$3},{A$1, A$2, A$3}
```

Clearly, the itemsets {`A$2, A$3`} and {`A$1, A$3`} do not make sense, since it is not meaningful to say that a group of genes share the second and third or first and third steps of 10% of the EST class in concern. Such meaningless itemsets can be excluded by considering only the itemsets that have `$1` item for each class and where all the numbers are in the consecutive order for each class.

In step 6, the entire data set is scanned again to identify the clusters to which each record belongs. Here, it is checked whether a record has the frequent-itemsets used in defining the clusters. This completes the clustering process and step 7 is concerned with presenting the results.

As mentioned earlier, the resulting clusters may be overlapping and may not be exhaustive. While classical clustering techniques ensure that the discovered clusters are not overlapping, in many real world situations overlapping clusters do exist and are useful. There have been some work on identifying overlapping clusters, particularly in the domain of bio-informatics as presented by Banerjee *et al.* [8]. The same can also be said about clusters which do not cover all the objects of the data set. Therefore, the clusters discovered are left as is.

4 Clustering of Zebrafish Genes Based on Their EST Makeup

4.1 Objective

We used our new clustering algorithm to identify clusters of genes based on the ESTs. First, it was expected to classify the ESTs in to a set of classes based on the tissue type and the development stage they are found. The possibility of employing the standard clustering techniques was also examined in this work. In particular, partitioning methods and hierarchical methods were explored. One main problem with these methods with regards to this data set is not presenting a clear meaning for the clusters identified. In addition, dealing with a large number of attributes (101, which is the number of EST classes) can also be problematic with some implementations of these techniques.

It was also intended to test the quality of clusters discovered using gene ontology data. These data can be obtained from the Gene Ontology Project [3]. It develops three ontologies to describe gene products in terms of their associated biological processes, cellular components and molecular functions. Each description in one of these three categories is given a unique number known as a Gene Ontology ID (GO-ID) and these GO-ID numbers can be used to describe gene products without ambiguity. The similarity between two genes in terms of their involvement in shared biological processes, presence in cellular components and molecular functions can be obtained by counting the number of common GO-IDs between the two genes.

4.2 Implementation

There are three independent data sets, as follows, and each of them required a significant amount of data preprocessing.

1. The data regarding the ESTs of zebrafish genes obtained from UniGene (the build of 16th July was used)
2. The data regarding ESTs of all species obtained from GenBank (the build of 3rd August was used)
3. The gene ontology data of zebrafish genes available from the Gene Ontology Project (the build of 15th August was used)

Regarding the second data set it was necessary to separate the ESTs of zebrafish from the set of all records. It was also necessary to extract only the GenBank accession number, which uniquely identifies each EST, and the tissue type and the development stage of the EST, which were to be used in identifying classes of ESTs from the records on zebrafish ESTs.

After extracting the accession number, tissue type and development stage of zebrafish ESTs, all the different combinations of tissue types and development stages were identified and each combination was recognized as an EST class. Each EST class was also given a unique class identification number to be used in the subsequent steps. Altogether 101 such classes were identified. Some of such EST classes are shown in Table 1. Note that null values were allowed in one field (tissue type or development stage).

From the first data set, the genes and their ESTs were extracted. Then the IDs of ESTs were replaced by their respective EST classes. At the end of this step, the records contained the EST classes of each gene. Then another program was used to identify the EST classes that had a percentage higher than the cut-off percentage and to list the items to be used in the clustering process based on the concept of frequency levels. The cut-off percentage used was 10%. At the end of this step, the data was ready to be clustered using frequent-itemsets.

Table 1. Examples for EST classes

EST Class ID	*Tissue Type*	*Development Stage*
1	myocardium, endocardium, vessel	Adult
2	embryonic	6 - 48 hours post fertilization
28	olfactory epithelium	null

In terms of Gene Ontology data, the first step was to separate the Gene ID and their GO-IDs from the other data. Identifying the Gene IDs required an additional step because the IDs used by Gene Ontology data were those defined by Zebrafish Information Network [7]. It was also necessary to do some additional processing regarding GO-IDs since in some records, their meaning was changed by another field which added qualifiers such as “not” and “contributes to”.

We used a publicly available implementation of the apriori algorithm [1] to discover frequent-itemsets. The minimum support level used for frequent-itemset mining was 5%. Clusters of genes were identified from the entire set of genes. In addition, the genes were divided in to groups based on the number of ESTs found in the genes, and clusters of genes were identified from the genes of each group. The rationale behind identify clusters in this manner is that genes having a similar number of ESTs might have some similarity in behavior. The ranges in the number of ESTs used in identifying groups of genes were based on similar ranges identified in UniGene. 10 such groups were identified, which were named G-0 to G-9. Genes of group G-0 are those listed as having no ESTs. These genes are mainly defined based on entire RNA sequences rather than on ESTs. Such genes do not play a major part here.

4.3 Results

Clusters of genes were identified from the entire set of genes as well as from each group of genes other than Group 0, which was excluded from the clustering process because genes of this group have no ESTs. Altogether 256 clusters were identified. An attempt was also made to measure their similarity using gene ontology data. A global similarity measure was calculated for all the genes with gene ontology data. Then a similarity measure of each cluster was compared with the global similarity measure. Table 2 summarizes the results.

We compared the similarity measure for genes of each cluster with the calculated global similarity measure, which is 13.6042%. The details of one cluster with a high similarity measure are presented below.

```
Cluster ID: G4-46 Group: G4 (5-8 ESTs) Common EST structure:
{Tissue Type = Embryo, Development Stage = 7 different stages}  - 10%
{Tissue Type = null, Development Stage = myoblast}              - 10%
Similarity Measure: 36.4416%
No of genes in the cluster: 345
No of genes with gene ontology data: 105
```

Table 2. Clusters identified

Group #	*# of clusters*	*# of interesting*[a] *clusters*	*Highest similarity measure*
All Genes	23	8	49.6712%
G-1	27	9	93.3756%
G-2	18	9	21.5147%
G-3	36	23	51.1322%
G-4	48	41	36.4416%
G-5	31	23	21.6830%
G-6	25	19	22.7529%
G-7	19	3	15.9517%
G-8	15	1	14.5355%
G-9	14	2	14.1515%

[a] These have a higher than normal similarity measure based on gene ontology data.

4.4 Discussion

The interestingness of the clusters of genes identified arises from the fact that they are active in the same tissue type and the same development stage. If the genes of a cluster that have gene ontology data exhibit a similarity in gene ontology data, it would probably indicate that these genes have a higher level of similarity. However, a higher similarity in gene ontology data can not be seen as directly leading to the conclusion that the genes of that cluster have similar behavior. It depends on several other factors also. The quality of the results is affected by the manner in which the data regarding ESTs are presented in GenBank. Currently, a consistent terminology is not used to describe tissue types and development stages and therefore two terms might in fact mean the same thing. A consistent terminology would lead to more accurate results.

This work demonstrates the use of the technique of clustering based on frequent-itemsets and frequency levels. Since it is a generic technique, this can be used in other domains as well. One obvious candidate for its application is the domain of text documents, where clustering based on frequent-itemsets have previously been used. The use of frequency levels together with frequent-itemsets would result in better clusters here. In addition, it can be applied in several other situations. One example would be to identify precincts or other geographical areas whose populations show a similarity based on factors such as ethnicity, religion or age.

5 Conclusion and Future Work

Our work shows that the technique of clustering based on frequent-itemsets and frequency levels is capable of identifying clusters in an effective manner. More research work is needed to ensure its applicability in different domains. It is important to show that this clustering technique can be generic rather than being restricted to a particular area. Such work would also be required to verify that the overlapping and non-exhaustive nature of the discovered clusters do not seriously hamper their usefulness. It is also necessary to verify the usefulness of the gene clusters discovered based on their EST makeup and to extract more information from them. The work presented in GEPIS [4] and EST miner [5] can be investigated more thoroughly as a part of such an exercise.

References

1. An implementation of the apriori algorithm. `http://www.ug.cs.usyd.edu.au/∼abright/`.
2. GenBank Database. `http://www.ncbi.nlm.nih.gov/Genbank/`.
3. Gene Ontology Project. `http://www.geneontology.org/index.shtml`.
4. GEPIS (Gene Expression Profiling in silico). `http://www.cgl.ucsf.edu/Research/genentech/gepis/gepis.html`.
5. Sorghum EST Clustering Analysis. `http://cggc.agtec.uga.edu/estMiner/estMiner.jsp`.
6. UniGene Database. `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene`.
7. ZFIN: The Zebrafish Information Network. `http://www.zfin.org`.
8. A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. J. Mooney. Model-based overlapping clustering. In *KDD*, pages 532–537, 2005.
9. F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *KDD*, pages 436–442, 2002.
10. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*, pages 440–444. Morgan Kaufmann Publishers, second edition, 2006.
11. Y. Li and S. M. Chung. Text document clustering based on frequent word sequences. In *CIKM*, pages 293–294, 2005.
12. H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *SIGMOD Conference*, pages 394–405, 2002.
13. X. Yan, H. Cheng, J. Han, and D. Xin. Summarizing itemset patterns: a profile-based approach. In *KDD*, pages 314–323, 2005.