KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2007
Cluj-Napoca (Romania), June 6–8, 2007, pp. 219–226

ON SOFTWARE ATTRIBUTES RELATIONSHIP USING A NEW FUZZY C-BIPARTITIONING METHOD

HORIA F. POP AND MILITON FRENŢIU

Abstract. A new data analysis method is introduced, fuzzy bipartitioning method, aimed at producing a set of fuzzy biclusters, that are pairs of fuzzy clusters of items and variables. The paper continues the data analysis of the dependence between software attributes performed in a former paper [14], as a case study to illustrate the importance of this new clustering method. The studied data set is formed by a number of projects written by second year students as a requirement in their curriculum.

1. Introduction

The main purpose of Software Metrics is to improve the Software development process [9]. Software Metrics are also useful to evaluate the quality of a software product [18]. And, as we show in this paper, Software Metrics are useful in education. The future programmer will respect an adequate programming methodology if he is taught to do so. The dependency between some software product attributes was discussed by many authors [1, 2, 25].

This study comes as a continuation of the previous work of both authors [10, 11, 12, 13, 14, 15].

2. The experiment

The study is based on 29 projects produced by second year undergraduate students as part of their requirements curriculum. These projects were analysed observing the attributes described in Table 1. Due to space constraints, the primary data is not given here, but it can be found in [14].

The attributes A10 and A11 were measured automatically by computer. All the others were estimated by postgraduate students. All metrics have the values in the interval [0, 10], where 0 stands for "very bad" (or not present at all), and 10 for "excellent". These values are the subjective evaluation of students. This

2000 *Mathematics Subject Classification.* 68N30.

Key words and phrases. Software metrics, Measurement, Biclustering, Fuzzy Clustering, Data Analysis Techniques, Education.

Attribute	Description	Attribute	Description
A1:	requirements description	A16:	readability
A2:	good specification	A17:	comprehensibility
A3:	function points	A18:	changeability (modifiability)
A4:	design clarity	A19:	structuredness
A5:	design correctness	A20:	testability
A6:	design completeness	A21:	reliability
A7:	design diagrams	A22:	efficiency
A8:	modules specification	A23:	extensibility
A9:	algorithms description	A24:	adaptability
A10:	lines of code	A25:	documentation clarity
A11:	no. of comments	A26:	documentation completeness
A12:	good use of comments	A27:	maintainability
A13:	good use of free lines	A28:	simplicity
A14:	indentation	A29:	quality
A15:	good names		

Table 1. Attributes description

subjectivity does not affect the attributes relationships, all values for a project being given by the same person. After all, "subjective measures are cheap and worth using" [7]. The definitions of the used attributes are inspired from and can be found in [9].

The attribute A12 refers to the documentation done by comments. It is not based on the number of comment lines of the programs. We may write as many comment lines as we like and sometimes the comments contradict the code, or do not reflect what the code does. The measure for this attribute takes in account if the specification of each module is reflected through comments, if the meaning of each variable and object is explained by comments, if the invariants and other important explanations are given by comments.

3. Overview of biclustering methods

Biclustering is a data mining technique that allows simultaneous clustering of rows and columns. The technique has originally been introduced in 1972 by J.A. Hartigan [16], and the term was first used in 2000 by Cheng and Church [4], in gene expression analysis. The concept of two-mode clustering, with the same meaning, has been introduced in 2004 by Van Mechelen, Bock and De Boeck [24].

Given a set of m rows in n columns, the biclustering algorithm generates biclusters, i.e. a subset of rows that exhibit similar behavior across a subset of columns, and vice-versa. Different biclustering algorithms have different definitions of bicluster.

Considering the relationships among the data, we identify:

- biclusters with constant values;
- biclusters with constant values on rows or columns;
- biclusters with coherent values;
- biclusters with coherent evolutions.

Considering the relationships between biclusters, the following bicluster structures may be obtained:

- exclusive row and column biclusters;
- non-overlapping biclusters with checkerboard structure;
- exclusive-rows biclusters;
- exclusive-columns biclusters;
- non-overlapping biclusters with tree structure;
- non-overlapping non-exclusive biclusters;
- overlapping biclusters with hierarchical structure;
- arbitrarily positioned overlapping biclusters.

For an excellent survey of biclustering algorithms see the survey paper of Madeira and Oliveira [17].

The paper [8] introduces a hybrid genetic fuzzy biclustering algorithm aimed at discovering value-coherent biclusters.

A different approach to biclustering, named cross-clustering, has been introduced in [6] and further studied and improved in [23].

A full-scale comparison of crisp and fuzzy cross-clustering and biclustering algorithms is beyond the purpose of this paper, and will be approached separately.

4. Fuzzy bipartitioning algorithm

The theory of fuzzy sets was introduced in 1965 by Lotfi A. Zadeh [26] as a natural generalization of the classical set concept. Let X be a data set, composed of n data items characterized by the values of s characteristics. A fuzzy set on X is a mapping $A : X \to [0, 1]$. The value $A(x)$ represents the membership degree of the data item $x \in X$ to the class A. The advantage of this approach is that it allows a data item x to be a member of more classes, with different membership degrees, according to certain similarity criteria.

Clustering algorithms based on fuzzy sets have proved their superiority due to their ability to deal with imprecise sets, imprecisely-defined boundaries, isolated points, and other delicate situations. The class of fuzzy clustering algorithms based on fuzzy objective functions [3] provides a large share of geometrical prototypes and combinations thereof, to be used according to the data substructure. On the other hand, the Fuzzy Divisive Hierarchical scheme [5, 19] provides an in-depth analysis of the data set, by deciding on the optimal subcluster cardinality and the optimal cluster substructure of the data set.

Let us consider the main point of biclustering algorihtms. They produce independent clusters formed by a selection of items and a selection of variables, such that the selected items are most similar by considering only the selected variables. Our aim is to generalize this approach in two different ways.

- we aim at producing fuzzy biclusters, not crisp ones;
- we aim at producing a set of biclusters, not only one.

Let us recall here that the fuzzy clustering algorithms of the FCM type use item-prototype dissimilarities based on distances between the item and the geometric prototype.

The main idea behind a fuzzy bicluster is that it is composed by two fuzzy sets: a fuzzy set of items and a fuzzy set of variables. The fuzzy set of variables actually define variables weights to be used in a weighted distance function when producing the fuzzy set of items. Similarly, the fuzzy set of items define the items weights to be used in a weighted distance function when producing the fuzzy set of variables.

This remark suggests an iterative procedure composed by two calls to various fuzzy clustering algorithms using adaptive metrics.

The method proposed in this paper is the following. We start by running a fuzzy horizontal variables clustering method [21]. In this way we will have a fuzzy partition formed by c fuzzy sets of variables.

For each of these c fuzzy sets, we run a fuzzy pointwise regression method[22, 20] (Fuzzy 1-Means), using a distance function weighted by the fuzzy membership degrees of the fuzzy variables set. This, actually, mean using a norm-induced distance with a diagonal norm matrix, with the fuzzy values on the diagonal.

We have now, a set of c fuzzy items sets, each one produced using one of the c fuzzy variables sets. For each of these c fuzzy items sets, we run a fuzzy pointwise regression method (Fuzzy 1-Means), using a distance function weighted by the fuzzy membership degrees of the fuzzy items set. This, actually, mean using a norm-induced distance with a diagonal norm matrix, with the fuzzy values on the diagonal.

At this point we have a set of c fuzzy variables sets, each one produced using one of the c fuzzy items sets. We are going to use these fuzzy variables sets in the same manner in order to produce a new set of fuzzy items sets.

This dual scheme continues until the two sets of fuzzy sets produced at an iteration are close enough to the fuzzy sets produced at the previous iteration.

Let us call this method *Fuzzy C-Bipartitioning (Regression) Method* and let us state it formally (the XT notation denotes the transpose of X).

```
Subalgorithm FuzzyCBipartitioningRegr (X, n, s, c, A, B) is
   Input:  X - data set with n items and s variables
           c - number of clusters to be produced, c>1
   Output: A - a set of c fuzzy sets of items, A[1], ..., A[c]
           B - a set of c fuzzy sets of variables, B[1], ..., B[c]

   Call FuzzyHorizontalClustering(XT, s, n, c, B)
   Repeat
      Let B’ := B;
      For i := 1 to c do
         Call FuzzyPointRegression(X, n, s, B’[i], A[i]);
      End for
      For i := 1 to c do
         Call FuzzyPointRegression(XT, s, n, A[i], B[i]);
      End for
   Until |B-B’| < eps
End subalg
```

A few remarks are in order. Firstly, this method does not produce fuzzy partitions. The sets A_i, $i = 1, \ldots, c$, do not form a fuzzy partition of X, and the same is valid for B_i, $i = 1, \ldots, c$. This is actually not a problem, but it may even be an advantage.

The spread of the fuzzy sets produced using fuzzy regression is controlled by the fuzzy regression method itself. We recall here that these fuzzy regression methods use an input parameter to denote the smallest fuzzy membership value to be assigned.

Finally, our fuzzy bipartitioniong method requires as input the number of fuzzy biclusters to be produced. While this may be considered a major drawback, it is actually not an issue. Let us recall that we are constructing biclusters and bipartitions because, on one side, we do have an idea about the fuzzy cluster substructure of a data set, and we need more info not on the number of relevant clusters, but the relationship between particular data clusters and the variables clusters that group the variables most important to explain these data clusters.

A variation of this algorithm would imply running Fuzzy Clustering instead of Fuzzy Regression at each step. However, this time an adaptive metric would be used on a per class basis. Due to the use of full clustering procedures, this algorithm will construct complete fuzzy partitions both for the data items, and for the variables.

We call this method *Fuzzy C-Bipartitioning (Clustering) Method.* Its formal description follows.

```
Subalgorithm FuzzyCBipartitioningClust (X, n, s, c, A, B) is
   Input:  X - data set with n items and s variables
           c - number of clusters to be produced, c>1
   Output: A - a partition of c fuzzy sets of items A[1],...,A[c]
           B - a partition of c fuzzy sets of variables B[1],...,B[c]

   Call FuzzyHorizontalClustering(XT, s, n, c, B)
   Repeat
      Let B' := B;
      Call FuzzyHorizontalAdaptiveClustering(X, n, s, B', A);
      Call FuzzyHorizontalAdaptiveClustering(XT, s, n, A, B);
   Until |B-B'| < eps
End subalg
```

5. Case study. Data analysis of software attributes

Based on the analyses performed in the previous papers, we have selected for our case study the use of the Fuzzy 5-Bipartitioning (Clustering), i.e. we aim at a partition of five classes. The bipartition obtained by defuzzyfication from the final fuzzy bipartition is available in Table 2. Due to space constraints, we are not giving here the table of fuzzy membership degrees to the five biclusters. This data is available upon request from the authors. The clustering error used for this case study is $eps = 10^{-5}$.

Class	Projects	Attributes
1	9 12 16 18 19 20 27	1 3 6 8 10 11 13 14 18 26 27 28 29
2	1 2 4 5 6 7 8 13 14 15 17 21 22 23 24 25 26 28 29	5 15 20 23 24
3	10	2
4	11	16 17
5	3	4 7 9 12 19 21 22 25

Table 2. The final fuzzy bipartition for five classes

A few comments with respect to the validity of these results. The first issue we should note is that we are discussing a fuzzy clustering method. As such, the whole set of advantages of fuzzy sets come as well with this method. There are many items with important fuzzy membership degrees to more than one class. See, for example, projects 16, 18, 20, and with a lesser extent, projects 8, 9, 12, 14, 19, 25, 27, all of these with respect to classes 1 and 2. As well, see attributes 29, 24, 20, 18, 15, 13, 5 (with respect to classes 1 and 2), attribute 7 (with respect to classes 1 and 5), attribute 25 (with respect to classes 2 and 5). These fuzzy membership degrees illustrate mixed behavior, not possible to be seen using crisp clustering methods.

A special mention for the qualitative attributes 18 to 29. They are distributed among three classes, many of them with important fuzzy membership degrees. The subjectivity of qualitative evaluations suggested by these attributes is indicated very clearly here by the fact that attribute 29, ‘quality’, has membership degrees of 0.556 and 0.436 to classes 1 and 2.

Similar correlations between numerical attributes, and between the design attributes, may be seen as well.

For project 10, the attribute 2 is considered very good, meanwhile the majority of the other attributes are very less important.

The clustering of projects 11 and 3 in separate classes is supported by an analysis of their attributes values. They show a lack of correlation between closely related attributes, issues that are normally not found among other projects. For example, project 3 has an overall quality grade of 5, even if all individual quality grades are 5, 6 or 7, the majority of them being greater than 5.

6. Further work and Concluding remarks

The method highlighted here is subject to be refined into a large family of fuzzy bipartitioning algorithms, each such method oriented towards a particular issue mentioned above. Further papers will, of course, have to consider:

(a) hierarchic fuzzy bipartitioning algorithms;
(b) refined algorithms producing sets of biclusters;
(c) refined algorithms producing actual fuzzy partitions;
(d) algorithms using different adaptive metrics, suitable for clusters of non-spherical shapes.

References

[1] Baecker R., Marcus A., *Design Principles for the Enhanced Presentation of Computer Program Source Text*, CHI’86 Proceedings, 51–58

[2] Basili V.R., Selby R.W., Hutchens D.H., *Experimentation in Software Engineering*, IEEE Transactions on Software Engineering, Vol. Se-12 (1986), no.7, 733–743

[3] Bezdek J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981

[4] Cheng Y., Church G.M., *Biclustering of expression data*, Proceedings of the 8-th International Conference on Intelligent Systems for Molecular Biology (2000), 93-103

[5] Dumitrescu D., *Hierarchical pattern classification*, Fuzzy Sets and Systems 28 (1988), 145–162

[6] Dumitrescu D., Pop H.F., Sârbu C., *Fuzzy hierarchical cross-classification of Greek muds*, Journal of Chemical Information and Computer Sciences 35 (1995), 851–857

[7] Dunsmore A., Roper M., *A Comparative Evaluation of Program Comprehension Measures*, EfoCS-35-2000, Department of Computer Science, University of Strathelgde, Glasgow, 2000

[8] Fei X., Lu S., Pop H.F., Liang L.R., *GFBA: A Genetic Fuzzy Biclustering Algorithm for Discovering Value-Coherent Biclusters*, Proceedings of the 2007 International Symposium on Bioinformatics Research and Applications, Georgia State University, Atlanta, Georgia, May 6-9, 2007

[9] Fenton N.E., *Software Metrics.* A Rigorous Approach, Int. Thompson Computer Press, London, 1995

[10] Frenţiu M., *On programming style – program correctness relation*, Studia Univ. Babes-Bolyai, Series Informatica 45, 2 (2000), 60–66

[11] Frenţiu M., *The Impact of Style on Program Comprehensibility*, "Babeş-Bolyai" University of Cluj-Napoca, Research Seminar on Computer Science, 2002, pp. 7–12

[12] Frenţiu M., Lazăr I., Pop H.F., *On individual projects in software engineering education*, Studia Universitatis Babes-Bolyai, Series Informatica 48, 2 (2003), 83–94

[13] Frenţiu M., Pop H.F., *A study of licence examination results using Fuzzy Clustering techniques*, Babeş-Bolyai University, Faculty of Mathematics and Computer Science, Research Seminars, Seminar on Computer Science, 2001, 99–106

[14] Frenţiu M., Pop H.F., *A study of dependence of software attributes using data analysis techniques*, Studia Universitatis Babes-Bolyai, Series Informatica 47, 2 (2002), 53–66

[15] Frenţiu M., Pop H.F., *Tracking mistakes in software measurement using fuzzy data analysis*, In The 4-th International Conference RoEduNet Romania (Sovata, Târgu-Mures, Romania, May 20-22 2005), pp. 150–157

[16] Hartigan J.A., *Direct clustering of a data matrix*, Journal of the American Statistical Association 67, 337 (1972), 123–129

[17] Madeira S.C., Oliveira A.L., *Biclustering Algorithms for Biological Data Analysis: A Survey*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1 (2004), 24–45

[18] McConnell S., *Software Quality at Top Speed, Software Development*, 1996, `http://www.construx.com/stevemcc`

[19] Pop H.F., *SAADI: Software for fuzzy clustering and related fields*, Studia Universitatis Babeş-Bolyai, Series Informatica 41, 1 (1996), 69–80

[20] Pop H.F., *Development of robust fuzzy regression techniques using a fuzzy clustering approach*, Pure Mathematics and Applications 14, 3 (2004), 221–232.

[21] Pop H.F., *Data analysis with fuzzy sets: a short survey*, Studia Universitatis "Babes-Bolyai", Series Informatica 49, 2 (2004), 111–122.

[22] Pop H.F., Sârbu C., *A new fuzzy regression algorithm*, Analytical Chemistry 68, 5 (1996), 771–778.

[23] Pop H.F., Sârbu C., *The fuzzy hierarchical cross-clustering algorithm. Improvements and Comparative study*, Journal of Chemical Information and Computer Sciences 37, 3 (1997), 510–516.

[24] Van Mechelen I., Bock H.H., De Boeck P., *Two-mode clustering methods: a structured overview*, Statistical Methods in Medical Research 13, 5 (2004), 363–94.

[25] Vessey I., Weber R., *Some Factors Affecting Program Repair Maintenance: An Empirical Study*, Comm. A.C.M., 26 (1983), no. 2, 128–134.

[26] Zadeh L.A., *Fuzzy sets*, Information and Control 8 (1965), 338–353.

Babeş-Bolyai University, Department of Computer Science, Faculty of Mathematics and Computer Science, 1 M. Kogălniceanu Street, 400084 Cluj-Napoca, Romania

E-mail address: `hfpop@cs.ubbcluj.ro`

E-mail address: `mfrentiu@cs.ubbcluj.ro`