Fuzzy Subtractive Clustering Based Indexing Approach for Software Components Classification

Sathit Nakkrasae[1,2,3]
[1]Department of Computer Technology, Faculty of Science, Ramkhamhaeng University, Bangkok, 10240, Thailand.
Khunsathit@hotmail.com

Peraphon Sophatsathit[2]
[2]Advanced Virtual and Intelligent Computing (AVIC) Center, Department of Mathematics, Faculty of Science, Chulalongkorn University, Bangkok, 10330, Thailand.
Peraphon.S@Chula.ac.th

William R. Edwards, Jr.[3]
[3]Center for Advanced Computer Studies (CACS), The University of Louisiana at Lafayette, Lafayette, LA 70504, U.S.A.
wre@cacs.louisiana.edu

Abstract

Software Engineering is not only a technical discipline of its own, but also a problem domain where technologies coming from other disciplines are relevant and can play important role. One important example is knowledge engineering [1], a term that used in a board sense to encompass artificial intelligence, computational intelligence, knowledge bases, data mining, and machine learning. Many of typical software development issues can benefit from these disciplines. For this reason, this paper will employ computational intelligence approach to classify software component repository into similar component cluster groups with the help of Fuzzy Subtractive Clustering algorithm. The center of each cluster will be used to construct the coarse grain classification indexing structure. Subsequent retrieval requirements of software component are compared with all the indexed cluster centers. Any software components belonging to the cluster partition whose center is closest to the required software component will be retrieved for subsequent participation in component selection at fine grain level. This approach not only is suitable for multidimensional data, but also automatically decides the correct model classification.

Keywords: Software component classification, knowledge engineering, neural networks, Fuzzy Subtractive Clustering.

1. Introduction

Reuse is a popular design methodology common to engineering discipline. It has two primary aspects: (a) cost reduction resulting from not design a new solution; and (b) increased confidence in the solution because of its successful reusing. For reuse to be an effective problem solving methodology, the designer must be able to reuse appropriate solutions, adapt a solution to fit the new problem, and evaluate the resulting solution. In software engineering, reuse is popularly applied in design domain. Owing to creativity and complexity of design paradigms, approaches, and the process itself, design reuse must, in many cases, be tailored to suit specific requirements. Moreover, automated software reuse support has been slow to emerge due to the difficulty in providing a useful design representation for software components. This design representation must be able to efficiently support component storage, retrieval, adaptation, and verification.

This paper presents software component matrix representation based on a well-defined software component specification [2] and an alternative software component classification and retrieval approach, utilizing computational intelligence on neural network [3].

The remaining of this paper is organized as follows. Section 2 discusses related papers on formal specification, and classification of software components. Other related topics are also incorporated. Matrix representation of software component based on structural, functional, and behavioral properties is presented in Section 3. Models and methodologies of software component classification in both coarse and fine grain are described in Section 4. Section 5 discusses the experiment and the results of component classification. Our final thought is discussed in Section 6.

2. Related Papers

A popular method for describing repositories of reusable software components is a faceted classification scheme [4]. Using this methodology, components are classified by a set of attribute-value pairs, or features. A domain expert, who is required to analyze the repository of software components and classify them according go predefined terms, performs the classification. The knowledge of domain expert is implicit in the classification. To provide a basis for similarity calculations, the terms that represent the set of possible values for a feature are often related by a conceptual distance graph [4]. The informality and imprecision of these classification schemes complicates the automation of the overall reuse process. Automation of the classification process requires reverse engineering from source code. The imprecision of the classification scheme does not support formal component verification; reasoning about identically classified components requires source code analysis.

The use of formal specifications to augment software reuse has been proposed to solve problems [2, 5, 6, 7, 8, 9, 10, 11]. There are many benefits to applying formal methods to software reuse. First, formal specifications provide an explicit representation of structure, function, and behavior of a software component free from many implementation details. This is valuable because structure, function, and behavior are the primary point of interest when determining reusability. Next, the expressiveness of formal specification languages allows precision beyond that of faceted classification. Equivalent specifications perform equivalent properties of software component (structural, functional, and behavioral). Finally, formal specifications and their associated formal system provide a basis for automated reasoning. A formal specification defines the structure, function, and behavior in terms of a domain model. A collection of axioms that define the data types and operations used in the system. Formal reasoning based on the domain model can be used logically verify the reusability of a software component.

This paper proposes a systematic approach for classifying software components in the form of machine learning through computational intelligence such as neural networks, fuzzy logic, and artificial intelligence. Some suggest Self-Organizing Map (SOM) [1, 12] for software component classification. However, there are limitations on this method:

- Its determination is not based on optimizing any model of process or its data;
- Prototype parameters may be severely affected by noise from data points and outliners. This is due to the fact that learning rates in SOM are computed as a function of the number of input presentations and node positions in the grid, while they are independent of the actual distance separating the input pattern from the cluster template;
- The size of the output lattice, the step size, and the size of the resonance neighborhood must vary empirically from one data set to another to achieve useful results; and
- It should not be employed in topology-preserving mapping when the dimension of the input space is larger than three [13].

Other popular classifying techniques are also take into account, such as Fuzzy C-Means clustering technique, which is a simple and straightforward approach but requires two predefined clusters where every data point membership depends on membership grade. It is clear from existing approaches that clustering technique is the fundamental building block of data classification. As such, we proposed Fuzzy Subtractive Clustering (FSC) technique [14] which is a fast one-pass algorithm for estimating the number of clusters and cluster centers in a set of data [15] to pre-process the software components. Once the software component groups are formed, classification process can proceed.

3. Software Component Representation

The proposed approach employs a formal specification [2] describing three properties of component, namely, structural, functional, and behavioral properties, free from most implementation details. These specifications are denoted in matrix form to support classification in the component repository. Subsequent retrieval of the desired component will utilize the same matrix to find the appropriate matching. As such, we will present a formulation of the classification matrix below.

Define software component X to be

$X = (S, F, B)$

where S denotes structural properties, F denotes functional properties, and B denotes behavioral properties. Each property is a list of the form

$S = \{ S_1, S_2, S_3, \ldots, S_m \}$,

$F = \{ F_1, F_2, F_3, \ldots, F_n \}$, and

$B = \{ B_1, B_2, B_3, \ldots, B_p \}$, respectively.

Each member of the list S, F, and B is also a list of the form

$S_i = \{ S_{i,1}, S_{i,2}, S_{i,3}, \ldots, S_{i,ui} \}$, $1 \le i \le m$ and $S_{i,j} \in D(S_i)$

$F_i = \{ F_{i,1}, F_{i,2}, F_{i,3}, \ldots, F_{i,vi} \}$, $1 \le i \le n$ and $F_{i,j} \in D(F_i)$

$B_i = \{ B_{i,1}, B_{i,2}, B_{i,3}, \ldots, B_{i,wi} \}$, $1 \le i \le p$ and $B_{i,j} \in D(B_i)$

and u_i, v_i, and w_i denote the number of members within S_i, F_i, and B_i, respectively. Each member is ordered from left to right, followed by the software component specification [2]. $D(S_i)$, $D(F_i)$, and $D(B_i)$ define separate equivalent classes (EC). For example, a system designer may wish to define a family of data objects to be stack-like, all belong to equivalent class of *LIFO*. This formulation entails a set of equivalent classes to be predefined within a component repository system by designers or developers.

Two preamble assumptions of our component repository stipulate that the number of elements in the set of equivalent classes for a given software component be finite, and that the number of each equivalent class in each property of the components be known. Denote the number of each structural, functional, and behavioral equivalent class properties by T_{Si}, T_{Fj}, and T_{Bk}, where $1 \le i \le m$, $1 \le j \le n$, and $1 \le k \le p$, respectively, we define property matrix representation as follows:

$$Col\_s = \text{Max}(T_{Si}, 1 \le i \le m),\ Row\_s = m$$
$$Col\_f = T_{F1},\ Row\_f = 1 + \Sigma^{n}_{i=2} T_{Fi}$$
$$Col\_b = T_{B1},\ Row\_b = 1 + \Sigma^{p}_{i=2} T_{Bi}$$

A software component matrix X can be written as follows:

$$C = (S_{Row\_s \times Col\_s},\ F_{Row\_f \times Col\_f},\ B_{Row\_b \times Col\_b})$$

For example, suppose $S = \{ S_1 \}$; $T_{S1} = 5$ (S_1 represents the component name in structural property of software component), $F = \{ F_1, F_2, F_3 \}$; $T_{F1} = 5$; $T_{F2} = 10$; $T_{F3} = 10$ (F_1, F_2, F_3 represent the function name, input, and output property of software component), $B = \{ B_1, B_2 \}$; $T_{B1} = 5$; $T_{B2} = 10$ (B_1, B_2 represent the behavioral name and action in behavioral property of software component). We further assume that C is made up of 3 functions and 4 behaviors as follows:

$C_{S1} = \{ S_{1,1} \}$

is the component structure name S_1 of C that contains $S_{1,1}$. Similarly, the function name F_1, input F_2, and output F_3 of C that represent the functional properties are positional arranged in matrix form as follows:

$C_{F1} = \{ F_{1,1} \}$
$C_{F2} = \{ F_{2,1}, F_{2,2}, F_{2,3} \}$
$C_{F3} = \{ F_{3,2}, F_{3,2} \}$

Similarly, function 2 becomes

$C_{F1} = \{ F_{1,2} \}$
$C_{F2} = \{ F_{2,3}, F_{2,10} \}$
$C_{F3} = \{ F_{3,2}, F_{3,10} \}$

and function 3

$C_{F1} = \{ F_{1,5} \}$
$C_{F2} = \{ F_{2,1}, F_{2,7} \}$
$C_{F3} = \{ F_{3,9}, F_{3.10} \}$

Similarly, the behavioral properties are denoted by

$C_{B1} = \{ B_{1,1} \}$
$C_{B2} = \{ B_{2,2}, B_{2,3}, B_{2,5} \}$

Behavior 2 becomes

$C_{B1} = \{ B_{1,3} \}$
$C_{B2} = \{ B_{2,3}, B_{2,3}, B_{2,6} \}$

Behavior 3 becomes

$C_{B1} = \{ B_{1,4} \}$
$C_{B2} = \{ B_{2,5}, B_{2,5}, B_{2,8}, B_{2,9} \}$

and behavior 4

$C_{B1} = \{ B_{1,5} \}$
$C_{B2} = \{ B_{2,2}, B_{2,3}, B_{2,7}, B_{2,10} \}$

The software component matrix is formed by concatenating individual i^{th} component vertically. Thus, $S = \{ C_{S1,n} \}$ where n denotes the column, $F = \{ C_{F1, n}$ concat $C_{F2, n}$ concat $C_{F3, n} \}$, and $B = \{ C_{B1, n}$ concat $C_{B2, n}$ concat $C_{B3, n}$ concat $C_{B4, n} \}$. Any missing property columns are denoted by zero. The resulting matrices S, F, and B are depicted in Figure 1.

Based on the above example, the algorithm for component matrix formulation proceeds as in Figure 2, where *Fnum* denotes the number of function in software component and *Bnum* denotes the number of behavior in the component. The final matrix becomes

$$X = (S, F, B)$$

This matrix will be transformed for use by subsequent proposed neural network computations.

4. Software Component Classification

Our approach for component classification is based on how software component is reused through the reuse model in order to establish a classification framework over the applicable component domain. We employed formal notations presented in [2] to represent component class, along with an example to demonstrate the applicability of the proposed framework. Various components are then grouped by coarse grain criteria (structure, function, and behavior) to form component clusters. We measured classification correctness by means of recall and precision techniques. If the result is satisfactorily, we proceed to fine grain classification to ensure proper reuse indicator assignment for the designated component or components are being retrieved from the cluster.

$$S = \begin{bmatrix} C_{S_1} = [\,1\ 0\ 0\ 0\ 0\,] \end{bmatrix}$$

$$B = \begin{bmatrix} C_{B_1} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 \end{bmatrix} \\ C_{B_2} = \begin{bmatrix} 0&0&0&0&0\\ 1&0&0&0&1\\ 1&0&2&0&1\\ 0&0&0&0&0\\ 1&0&0&2&0\\ 0&0&1&0&0\\ 0&0&0&0&1\\ 0&0&0&1&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{bmatrix} \end{bmatrix} \qquad F = \begin{bmatrix} C_{F_1} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \end{bmatrix} \\ C_{F_2} = \begin{bmatrix} 1&0&0&0&1\\ 1&0&0&0&0\\ 1&1&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&1\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&1&0&0&0 \end{bmatrix} \\ C_{F_3} = \begin{bmatrix} 0&0&0&0&0\\ 2&1&0&0&0\\ 0&1&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&1\\ 0&0&0&0&1 \end{bmatrix} \end{bmatrix}$$

Figure 1. Software Component Matrix Representation

```
First: S Row_s×Col_s

    For i = 1 to m
        For j = 1 to T_Si
            S (i,j) = l (S_i has l terms in equivalent class j)

Second: F Row_f×Col_f

    For i = 1 to Fnum
    Begin
        j = Eq_class_number(function(i))
        F(1,j) = 1
        For k = 2 to Row_f
            F(k,j) = m (F_j has m terms in equivalent class k)
    End for I

Third: B Row_b×Col_b

    For i = 1 to Bnum
    Begin
        j = Eq_class_number(function(i))
        B(1,j) = 1
        For k = 2 to Row_b
            B(k,j) = p (B_j has p terms in equivalent class k)
    End for I
```

Figure 2. Matrix Calculation Algorithm

4.1 Software Reuse Model

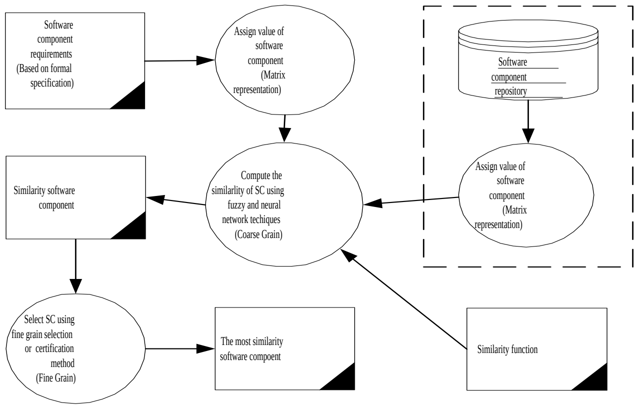
The software reuse model encompasses a repository which stores formal specifications of software components and retrieval mechanisms to facilitate component check-in/check-out during the development process. The underlying principle of the proposed classification scheme relies on component similarity comparison that is derived from a user-defined classification function. This offers a quantitative technique to enumerate the component suitability in coarse grain level. Assessment begins by representing software components in matrix form. This permits quantitative evaluation of the requirement specification of the designated component and the components stored in the repository, based on the requirement specification of the cluster component. Subsequent classification process will sort out the closest matched component for reuse purpose. Detail on how classification and retrieval are described in the next section. Evaluation is performed using FSC algorithm to arrive at a similarity value. This process is depicted in Figure 3.

4.2 Coarse Grain Level Software Component Classification Techniques

The three properties used to classify software components are component structure, function, and behavior. The component structure is made up of a component name, a subcomponent name, a class name, a signature, and an interaction name. The component function consists of a function name, input parameters, local variables, output parameters, and pre/post expressions. The component behavior is composed of a behavior name, a state name, and an action name. These properties are represented in matrix form described in Section 3.

The classification process starts by dividing software components into groups using FSC algorithm. These clusters are used to construct the index structure as shown in Figure 4. Searching for the closest match between the specified requirement and those in the index yields optimal software component retrieval. When a match is found, all components belonging to that cluster are retrieved. In general, more than one match may result. A fine grain certification step is required to ensure the best candidate being retrieved. Details on how certification proceeds will be postponed till Section 4.3.

Two measures of software component retrieval performance used in this paper are recall and precision [16]. Recall is the ratio of the number of relevant items retrieved to the total number of relevant items in the repository. High recall indicates that relatively few relevant software components were overlooked. Precision is the ratio of the number relevant items retrieved to the total number of items retrieved. High recall indicates that relatively few relevant software components were overlooked. Precision is the



Software component requirements (Based on formal specification)
Assign value of software component (Matrix representation)
Software component repository
Similarity software component
Compute the similarlity of SC using fuzzy and neural network techiques (Coarse Grain)
Assign value of software component (Matrix representation)
Select SC using fine grain selection or certification method (Fine Grain)
The most similarity software compoent
Similarity function

Figure 3. Software reuse model

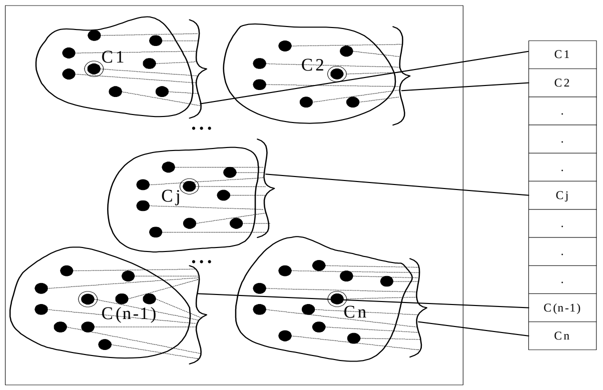


Figure 4. The n cluster partitions generated by FSC algorithm. The black dots and double line black dots denote software components and cluster centers, respectively

ratio of the number relevant items retrieved to the total number of items retrieved. High precision means that relatively few irrelevant software components were retrieved. In general, there is tradeoff between precision and retrieval. The goal is to find a practical balance between the two. The relevance condition is fundamental to the evaluation of a retrieval system.

It was also informative to observe the number of software components retrieved by the system. This number can help estimate the load that would be placed on the designer to interpret the results of a query in an interactive system, or similarly, the search space that would be faced by an adaptation system when considering software component compositions.

Given a set of a priori clusters $C = \{c\}_1^n$ and the calculated FSC clusters, $C' = \{c'\}_1^m$, the performance measures of FSC is defined as follows [17]:
Recall = Number of target software components retrieved / Number of target software components

$$= \sum_{c_i \in C \wedge c'_j \in C'} \frac{c_i \cap c'_j}{\# c_i} \qquad Eq\ (1)$$

Precision = Number of target software components retrieved / Number of software components retrieved

$$= \sum_{c_i \in C \wedge c'_j \in C'} \frac{c_i \cap c'_j}{\# c'_j} \qquad Eq\ (2)$$

where $\#c_i$ denoted the number of elements on cluster c_i and $0 \le$ *recall*, *precision* ≤ 1. Based on the above definitions, recall expresses the ratio of the target repository objects being actually retrieved out of all the expected target repository objects, whereas precision indicates the ratio of target repository objects to the retrieved set. For example, there are 10 repository objects and 4 of them are pre-specified as target repository objects. Given a query retrieving 5 objects and 3 out of those five objects are target objects. In this case, recall is 0.75 and precision is 0.6. The higher the recall and precision get the more accurate the method for retrieval becomes. We can calculate the accuracy for each FSC clusters based on the information pertaining to their natural clusters. The response time of the system was measured to determine the practicality of the method. For each measured quality, the minimum and median were calculated from every scenario in the experiment, which will be discussed in Section 5.

4.3 Fine Grain Level Software Component Selection Technique

In this level, we will try to find the most suitable software component for reuse. The degree of significance defined by user will be used as the selection criteria. The following notations will be given:

- $Ø_s$, $Ø_f$, and $Ø_h$ are the degree of significance of structural, functional, and behavioral properties, respectively, satisfying $0 \le Ø_s, Ø_f, Ø_h \le 1$ and $Ø_s + Ø_f + Ø_h = 1$. The degree of significance depends on system environment under which developers can define in accordance with the underlying system;
- N_r is the number of retrieved software components from the cluster whose center is closest to the required software component;
- X_i is the i^{th} retrieved software component in the component matrix described in Section 3, i.e., $X_i = (S, F, B)$ where $1 \le i \le N_r$;
- X_r is the component requirements; and
- *SC* is the most suitable software component which can be determined as follows:

$$SC = X_{reuse} \qquad Eq\ (3)$$

where the value of reuse can be computed from

$$reuse = \arg \min_{1 \le i \le N_r} \Big(\sum_{p=S,F,B} \phi_p \, || Xp_r - Xp_i || \Big) \qquad Eq\ (4)$$

5. Experiment

5.1 Data Collection

One hundred software component specification data were generated by uniform distribution generator. The data were arranged in matrix form suitable for the proposed algorithm described in Section 3. Components were classified according to their structural, functional, and behavioral properties, which in turn, were grouped into appropriate equivalent classes. In so doing, each component data vector encompassed 1320 dimensions.

The data set was divided into two groups, namely, 50 training set and 50 test set. Each data vector was normalized within [0,1] according to

$$v_{new} = \frac{v_{old} - v_{\min}}{v_{\max} - v_{\min}} \qquad Eq\ (5)$$

where v_{new} is the new value of the designated variable for that data point, v_{old} is the old value of the data point, v_{min} is the minimum value of the variable from all data points, and v_{max} is the maximum value of the variable from all data points.

5.2 Cluster Center Detection

We selected FSC approach to determine cluster centers using parameter $r_a = 14$. This value is the maximum distance between any two points within the same cluster, yet less than the distance between any two points from different clusters where each point belongs. The multiplier $Sqsh = 1.25$ is the default squash factor value of MATLAB 5.3.

The criteria for cluster center consideration are based on acceptance and rejection ratios. Acceptance ratio is the fractions of the potential first cluster center above which another data point will be accepted. Rejection ratio is the condition to reject a data point to be a cluster center, obtained from the fraction of the potential first cluster center below which a data point will be rejected as a cluster center. We chose 0.5 as the accepted ratio (default value from MATLAB version 5.3) for the first cluster center. We chose the rejection ratio (η) between 0.15-0.5 to derive other cluster centers. The resulting rejection ratios from various cluster centers were used to compare and evaluate the component classification. The procedure for grouping 50 data point clusters $\{X_1, X_2, X_3, \ldots, X_{n=50}\}$ in the training set is described below.

1) Compute the initial potential value for each data point (x_i)

$$P_i = \sum_{j=1}^{n} e^{-\alpha || x_i - x_j ||^2} \qquad Eq\ (6)$$

where $\alpha = 4/r_a^2$

$\| . \|$ is the Euclidean distance

r_a is a positive constant representing a normalized neighborhood data radius.

Any point falls outside this encircling region will have little influence to the potential point. The point with the highest potential value is selected as the first cluster center. This tentatively define the first cluster center.

2) A point is qualified as the first center if its potential value ($P^{(1)}$) is equal to the maximum of initial potential value ($P^{(1)*}$)

$$P^{(1)*} = \max_i (P^{(1)}(x_i)) \qquad Eq\ (7)$$

3) Define a threshold δ as the decision to continue or stop the cluster center search. This process will continue if the current maximum potential remains greater than δ.

δ = *(reject ratio)$\times$ (potential value of the first cluster center)*

where the rejection ratio (η) used in this work is 0.15-0.5, and $P^{(1)*}$ is the potential value of the first cluster center.

4) Remove the previous cluster center from further consideration.

5) Revise the potential value of the remaining points according to the equation

$$P_i = P_i - P_k^* e^{-\beta \| x_i - x_k^* \|^2} \qquad Eq\ (8)$$

where x_k^* is the point of the kth cluster center, P_k^* is its potential value, and $\beta = 4/1.25$ (*sash* * r_a).

6) For the point having the maximum potential value, calculate the acceptance ratio. If this value is greater than the predefined constant (0.5), the point is accepted to be the next cluster center. Otherwise, compute the rejection ratio. If the rejection ratio is greater than the predefined threshold (η = 0.15-0.5), this point is accepted. .

This procedure is repeated to generate the cluster centers until the maximum potential value in the current iteration is equal to or less than the threshold δ. After applying the subtractive clustering, we get different cluster center numbers from 50 training patterns depending on different rejection ratios. We used these different cluster center numbers to compare and evaluate software classification results as show in Table 1-5.

5.3 Evaluation

From the 100 vector data participated in the experiment, we regulated the rejection ratio in the range of 0.15-0.5 to avoid high rejection rate, whereby yielding too many unclassified or misclassification of the above data.

We assessed the accuracy of FSC algorithm by measuring recall and precision performance. The derived centers were anticipated to correctly classify software components in repository into each group of its predefined equivalent classes. From the 50 training data sets with predefined 10 equivalent classes, we conducted 5 trials using the remaining 50 test data sets with different rejection ratio (η) groups (0.15-0.20, 0.25, 0.30, and 0.35-0.50) and used the derived cluster centers from each trial to calculate its recall and precision performance. The results can be interpreted as follows. Based on 0.15-0.20, 0.25, 0.30, and 0.35-0.50 rejection ratio (η) groups, the values of cluster centers so derived are 18, 15, 11, and 10. Table 1, 2, 3, and 4 show the recall and precision results of 0.15-0.20, 0.25, 0.30, and 0.35-0.50 rejection ratio (η) groups, respectively.

Table 5 shows the comparative results of all 4 rejection groups obtaining from FSC classification of software component being quite satisfactory. Note that recall performance suffers a slight drop due to the decreasing of rejection ratio (η). There was, however, no fault classification in each equivalent class since the centers were closely located to their corresponding component groups. As such, retrieval was accomplished with relatively few attempts. The high number of centers selected from the 10 equivalent classes having 0.15-0.20, 0.25, and 0.30 rejection ratio (η) implied that there were more than one center in each equivalent class representing subgroup classification within each cluster, whereby yielding 98-100 % accuracy.

The above results were cross-validate with conventional statistical approach based on the same data sets. For well-dispersed clusters of data, the proposed approach performed equally well as the conventional statistical approach. This is because the cluster center in the conventional statistical approach was derived from the member average dispersion, which was the same as that obtained from the proposed approach. As data point began to overlap across cluster boundaries, the proposed approach out-performed the conventional statistical approach in that some clusters could end up having more than one center representing smaller groups within the same cluster. More accurate classification was thus resulted.

5.4 Example of Fine Grain Selection

We employed cluster centers from Table 1 to build a cluster index to enhance component searching and selection. Suppose X denotes a software component requirement matrix, which matches cluster center *SC19*, this cluster encompasses *SC16*, *SC17*, *SC18*, *SC19*, and *SC20*. Table 6 shows the results of most suitable software component selection from Equation 3 having different degrees of significance on structural, functional, and behavioral properties. Consequently, software developers can selectively apply the retrieved component or components that fit their respective purpose and application.

Table 1. Recall and Precision performance of 0.15-0.20 rejection ratio (η) value

Data Number Selected as Cluster Center	Software Component Relevant	tware Component RetrievedSof	Recall	Precision
1SC	5CS,4CS,3CS,2CS,1SC	1SC	0.20	1.00
3SC	5CS,4CS,3CS,2CS,1SC	5CS,4CS,3CS,2SC	0.80	1.00
6SC	10CS,9CS,8CS,7CS,6SC	7CS,6SC	0.40	1.00
8SC	10CS,9CS,8CS,7CS,6SC	10CS,9CS,8SC	0.60	1.00
11SC	15CS,14CS,13CS,12CS,11SC	11SC	0.20	1.00
12SC	15CS,14CS,13CS,12CS,11SC	12SC	0.20	1.00
15SC	15CS,14CS,13CS,12CS,11SC	15CS,14CS,13SC	0.60	1.00
19SC	20CS,19CS,18CS,17CS,16SC	20CS,19CS,18CS,17CS,16SC	1.00	1.00
22SC	25CS,24CS,23CS,22CS,21SC	25CS,24CS,23CS,22CS,21SC	1.00	1.00
26SC	30CS,29CS,28CS,27CS,26SC	26SC	0.20	1.00
28SC	30CS,29CS,28CS,27CS,26SC	30CS,29CS,28CS,27SC	0.80	1.00
31SC	35CS,34CS,33CS,32CS,31SC	31SC	0.20	1.00
33SC	35CS,34CS,33CS,32CS,31SC	35CS,34CS,33CS,32SC	0.80	1.00
37SC	40CS,39CS,38CS,37CS,36SC	37CS,36SC	0.40	1.00
38SC	40CS,39CS,38CS,37CS,36SC	40CS,39CS,38SC	0.60	1.00
43SC	45CS,44CS,43CS,42CS,41SC	45CS,44CS,43CS,42CS,41SC	1.00	1.00
46SC	50CS,49CS,48CS,47CS,46SC	46SC	0.20	1.00
50SC	50CS,49CS,48CS,47CS,46SC	50CS,49CS,48CS,47CS,46SC	0.80	1.00
	Average		0.56	1.00

Table 2. Recall and Precision performance of 0.25 rejection ratio (η) value

Data Number Selected as Cluster Center	Software Component Relevant	Software Component Retrieved	Recall	Precision
1SC	5CS,4CS,3CS,2CS,1SC	1SC	0.20	1.00
3SC	5CS,4CS,3CS,2CS,1SC	5CS,4CS,3CS,2SC	0.80	1.00
6SC	10CS,9CS,8CS,7CS,6SC	7CS,6SC	0.40	1.00
8SC	10CS,9CS,8CS,7CS,6SC	10CS,9CS,8SC	0.60	1.00
11SC	15CS,14CS,13CS,12CS,11SC	12CS,11SC	0.40	1.00
15SC	15CS,14CS,13CS,12CS,11SC	15CS,14CS,13SC	0.60	1.00
19SC	20CS,19CS,18CS,17CS,16SC	20CS,19CS,18CS,17CS,16SC	1.00	1.00
22SC	25CS,24CS,23CS,22CS,21SC	25CS,24CS,23CS,22CS,21SC	1.00	1.00
26SC	30CS,29CS,28CS,27CS,26SC	30CS,29CS,28CS,27CS,26SC	1.00	1.00
28SC	30CS,29CS,28CS,27CS,26SC	30CS,29CS,28CS,27CS,26SC	1.00	1.00
33SC	35CS,34CS,33CS,32CS,31SC	35CS,34CS,33CS,32CS,31SC	1.00	1.00
37SC	40CS,39CS,38CS,37CS,36SC	36SC	0.20	1.00
38SC	40CS,39CS,38CS,37CS,36SC	40CS,39CS,38CS,37SC	0.80	1.00
43SC	45CS,44CS,43CS,42CS,41SC	45CS,44CS,43CS,42CS,41SC	1.00	1.00
50SC	50CS,49CS,48CS,47CS,46SC	50CS,49CS,48CS,47SC,46SC	1.00	1.00
	Average		0.73	1.00

Table 3. Recall and Precision performance of 0.30 rejection ratio (η) value

Data Number Selected as Cluster Center	Software Component Relevant	Software Component Retrieved	Recall	Precision
3SC	5CS,4CS,3CS,2CS,1CS	5CS,4CS,3CS,2CS,1SC	1.00	1.00
6SC	10CS,9CS,8CS,7CS,6SC	6SC	0.20	1.00
8SC	10CS,9CS,8CS,7CS,6SC	10CS,9CS,8SC	0.60	1.00
15SC	15CS,14CS,13CS,12CS,11SC	15CS,14CS,13CS,12CS,11SC	1.00	1.00
19SC	20CS,19CS,18CS,17CS,16SC	20CS,19CS,18CS,17CS,16SC	1.00	1.00
22SC	25CS,24CS,23CS,22CS,21SC	25CS,24CS,23CS,22CS,21SC	1.00	1.00
28SC	30CS,29CS,28CS,27CS,26SC	30CS,29CS ,28CS,27CS,26CS,7SC	1.00	0.83
33SC	35CS,34CS,33CS,32CS,31SC	35CS,34CS,33CS,32CS,31SC	1.00	1.00
38SC	40CS,39CS,38CS,37CS,36SC	40CS,39CS,38CS,37CS,36SC	1.00	1.00
43SC	45CS,44CS,43CS,42CS,41SC	45CS,44CS,43CS,42CS,41SC	1.00	1.00
50SC	50CS,49CS,48CS,47CS,46SC	50CS,49CS,48CS,47CS,46SC	1.00	1.00
	Average		0.89	0.98

Table 4. Recall and Precision performance of 0.35-0.50 rejection ratio (η) value

Data Number Selected as Cluster Center	Software Component Relevant	Software Component Retrieved	Recall	Precision
3SC	5CS,4CS,3CS,2CS,1SC	5CS,4CS,3CS,2CS,1SC	1.00	1.00
8SC	10CS,9CS,8CS,7CS,6SC	10CS,9CS,8CS,6SC	0.80	1.00
15SC	15SC,14CS,13CS,12CS,11SC	15CS,14CS,13CS,12CS,11SC	1.00	1.00
19SC	20CS,19CS,18CS,17CS,16SC	20CS,19CS,18CS,17CS,16SC	1.00	1.00
22SC	25CS,24CS,23CS,22CS,21SC	25CS,24CS,23CS,22CS,21SC	1.00	1.00
28SC	30CS,29CS,28CS,27CS,26SC	30CS,29CS ,28CS,27CS,26CS,7SC	1.00	0.83
33SC	35CS,34CS,33CS,32CS,31SC	35CS,34CS,33CS,32CS,31SC	1.00	1.00
38SC	40CS,39CS,38CS,37CS,36SC	40CS,39CS,38CS,37CS,36SC	1.00	1.00
43SC	45CS,44CS,43CS,42CS,41SC	45CS,44CS,43CS,42CS,41SC	1.00	1.00
50SC	50CS,49CS,48CS,47CS,46SC	50CS,49CS,48CS,47CS,46SC	1.00	1.00
	rageAve		0.98	0.98

Table 5. Recall and Precision performance comparison

Rejection Ratio	Number of Center Selected	Recall	Precision
0.15-0.20	18	0.56	1.00
0.25	15	0.73	1.00
0.30	11	0.89	0.98
0.35-0.50	10	0.98	0.98

Table 6. Software Component Selection with Different Degree of Significance

Degree of Significance			Value Result of Each Software Component					The most Suitable Software Component for Reuse
Structural	Functional	Behavioral	16SC	17SC	18SC	19SC	20SC	
0.8	0.1	0.1	4.2557	4.2310	3.9437	4.0837	4.3289	18SC
0.1	0.8	0.1	7.2205	7.3283	6.8682	6.8349	7.6845	19SC
0.1	0.1	0.8	7.5370	7.2779	7.2057	7.3666	7.4824	18SC
0.3	0.3	0.3	5.7028	5.6483	5.4049	5.4846	5.8461	18SC

6. Conclusion

We have proposed two computational intelligent approaches to classify software components for effective archival and retrieval purposes, namely, fuzzy subtractive clustering algorithm and neural network technique. Component specifications are represented in matrix form to quantitatively organize these software artifacts for subsequent applications. Components were indexed based on the cluster centers so obtained. As such, subsequent reference and retrieval could be carried out efficiently through this indexing mechanism. We also conducted an experiment to assess the validity of the proposed approach, which turned out to be quite satisfactory.

We envision in our future work concerning software certification process to benefit from this rigorous formulation that will eventually be incorporated as part of the machine learning research endeavor. As a consequence, pervasive use of software components in the same manner as their hardware counterparts, as well as the ultimate COTS application, can be realized.

Acknowledgements

We would like to express our appreciate Prof. Dr. Chidchanok Lursinsap, Ms. Atchara Mahaweerawat, Mr. Khamron Sunat, Mrs. Wiphada Wettayaprasit of AVIC at Chulalongkorn University for their insight and constructive suggestions of our paper.

References

[1] W. Pedrycz, "Computational Intelligence as an Emerging Paradigm of Software Engineering", in *Proc. the Fourteenth International Conference on Software Engineering and Knowledge Engineering*, 2002, pp.7-14.

[2] S. Nakkrasae, and P. Sophatsathit, "Formal Approach for Specification and Classification of Software Components", in *Proc. the Fourteenth International Conference on Software Engineering and Knowledge Engineering*, 2002, pp.773-780.

[3] S. Haykin, "*Neural Network*", Prentice Hall, pp.256-312, 1999.

[4] E. Ostertag, J. Hendler, R. P. Diaz, and C. Braun, "Computing similarity in a reuse library system: An AI Base Approach", *ACM Trans. Software Engineering and Methodology*, pp. 205-228, 1992.

[5] A. M. Zaremski, and J. M. Wing, "Specification matching software components", in *the 3rd ACM SIGSOFT Symposium on the Foundations of Software Engineering*, 1995.

[6] D. E. Perry, and S. S. Popovitch, "In quire: Predicate-Based Use and Reuse", in *Proc. the 8th Knowledge-Based Software Engineering Conference*, 1993, pp. 144-151.

[7] S. A. Ehikioya, "A formal model for the reuse of software specifications", in *IEEE Canadian Conference on Electrical and Computer Engineering*, Volume: 1, 1999, pp. 283-288.

[8] J. J. Jeng, and B. H. C. Cheng, "Using formal methods to construct a software library", in *Proc. 4th European software Engineering Conference, Lecture Notes in Computer Science*, 1993, pp.397-417.

[9] J. J. Jeng, and B. H. C. Cheng, "A formal approach to using more general components", in *Proc. the 9th Knowledge-Based Software Engineering Conference*, 1994, pp. 90-97.

[10] P. Devanbu, R. J. Brachman, P. G. Selfridge, and B. W. Ballard, "LaSSIE: A knowledge-base software assistant", *Communications of the ACM*, pp. 34-49, 1991.

[11] R. S. Pressman, "*Software Engineering, A Practitioner's Approach*", 4th Editon, New York:McGraw-Hill, 1997.

[12] S. M. Charters, C. Knight, N. Thomas and M. Munro, "Visualization for informed decision making; Form code to components", in *Proc. the Fourteenth International Conference on Software Engineering and Knowledge Engineering: SEKE'02*, 2002, pp.765-772.

[13] A. Baraldi, and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition Part 2", *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics*, Vol. 29 No.6, 1999.

[14] S. Chiu, "Method and Software for Extracting Fuzzy Classification Rules by Subtractive Clustering", in *Fuzzy Information Proceeding Society, Biennial Conference of the North American*, 1996, pp. 461-465.

[15] P. Eklund, L. Kallin and T. Riissanen, "*Fuzzy Systems*", Lecture notes prepared for courses at Department of computing Science at Aumea University, Sweden, February, 2000.

[16] H. Mili, F. Mili, and A. Mili, "Reusing software: Issues and research directions", *IEEE Trans. Software Eng.*, pp. 528-562, 1995.

[17] I. King, and T. K. Lau, "Performance analysis of clustering algorithm for information retrieval in image databases", *International Joint Conference on IEEE World Congress on Computational Intelligence*, 1998, pp.932-937.

Sathit Nakkrasae is a lecturer in the department of Computer Technology at Ramkhamhaeng University and a Ph.D. student at Chulalongkorn University Mr. Nakkrasae's research focus is software engineering and knowledge engineering. He has identified specified, and classified software components using formal method and neural network. Mr. Nakkrasae has given presentations at international conferences and workshops. He has reviewed papers for IJSEKE (International Journal on Software Engineering and Knowledge Engineering) and has been invited to review for IEEE Trans Fuzzy Systems.

Peraphon Sophatsathit received the BS in industrial engineering from Chulalongkorn University, Thailand, the MS degrees in industrial engineering and computer science from the University of Texas at Arlington, and the Ph.D. degree in computer science from the Arizona State University. He is an assistant professor of computer science at Chulalongkorn University. His research interests include software engineering environments, design, and construction.

William R. Edwards is an Associate Professor in the Center for Advanced Computer Studies of the University of Louisiana at Lafayette, Louisiana. He received his Ph.D. in Computer Science from the University of Kansas in 1973. His research interests are in theory of computation and languages, information theory, and their application to software engineering. He is a member of the ACM and IEEE.