論文 98-23-4-27

The Database Construction of a Classification System Using an Optimal Cluster Analysis Model

Hyun-Sook Rhee*, *Regular Member*

최적 클러스터 분석 모델을 이용한 분류시스템의 데이타베이스 구축

正會員 이 현 숙*

ABSTRACT

Classification techniques are often an important component of intelligent systems and are used for both data preprocessing and decision making. In the design of a classification system, the labled samples must be given to provide *a priori* information for the classification. Moreover, the number of classes to be categorized must be known *a priori*. In this paper, we introduce an optimal cluster analysis model for carrying out fuzzy clustering without *a priori* information, called OFCAM. In OFCAM, an unsupervised learning algorithm for fuzzy clustering and a cluster validity strategy are integrated. Based on the information obtained by OFCAM, the database of a classification system, called PCSDB, is constructed. Then, PCSDB can be effectively used in the decision process of the system.

요 약

데이타의 분류기법은 공장자동화나 로보틱스 분야에서 사용되는 지능시스템의 중요한 기능이다. 일반적으로 이러한 분류시스템을 설계하고자 할 때, 준비된 데이타는 레이블링 되어야 하고, 분류하고자하는 클래스의 수도 설정되어야한다. 본 연구에서는 이러한 사전 정보없이 분류 시스템을 설계하고자 최적 클러스터 분석 모델, OFCAM을 제안한다. 이때 사용되는 최적 클러스터 분석 모델은 데이타의 구조에 대한 사전정보 없이, 주어진 데이타의 최적 클러스터의 수와 클러스터 중심점 및 각 데이타에 대한 소속 정보를 구해준다. 이를 위하여 OFCAM에서는 목적함수를 가지는 비교사 학습신경망과 클러스터 타당성 전략을 결합하고 있다. OFCAM의 결과를 바탕으로 분류시스템의 데이터베이스, PCSDB가 구축되며 이는 결정 모듈에서 쉽게 활용될 수 있음을 보인다. 이와같은 방법은 하나의 데이타베이스 안에서 필요한 테이블만을 첨가하므로 독립적으로 여러 응용의 분류문제를 다룰 수 있다.

*동양공업전문대학 전산경영기술공학부
論文番號:96277-0830
接受日字:1996年 8月 30日

Ⅰ. Introduction

Classification techniques are often an important component of intelligent systems and are used for both data preprocessing and decision making[1]. Thus, it has practical applications in a variety of fields, including pattern recognition and artificial intelligence, statistics, cognitive psychology, vision analysis, and medicine. In the design of a classification system, the labled samples must be given to provide *a priori* information for the classification. Moreover, the number of classes to be categorized must be known *a priori*. In this paper, we introduce an optimal cluster analysis model for carrying out fuzzy clustering without *a priori* information, called OFCAM(Optimal Fuzzy Cluster Analysis Model). In OFCAM, we devise the batch learning algorithm for the unsupervised fuzzy cluster analysis[2]. This algorithm is iterated for increasing number of clusters in given interval, computing the value of validity measure, which was defined in [3]. And this model uses a validity strategy[3], which selects an optimal number of clusters presented in the data using the relative values and normalized value of the validity measure. Based on the informatio1n obtained by OFCAM, the database of a classification system, called PCSDB, is constructed. Then, PCSDB can be effectively used in the decision process of the system and be systematically maintained in the classification system.

Ⅱ. Optimal Fuzzy Cluster Analysis Model

In this section, we are going to introduce an optimal cluster analysis model for carrying out fuzzy clustering without *a priori* information, called OFCAM. In order to find c cluster centers and membership information for any given data sets, we devise the batch learning algorithm, called FCGD-B[2]. FCGD-B has advantages over fuzzy c-means algorithm[4] and self-organizing maps[5] by integrating an optimization function into unsupervised learning networks. Moreover, the learning rules for FCGD-B are a result of formal derivation based on gradient descent method of a fuzzy objective function. FCGD-B does not suffer from several problems of conventional learning networks devised on the basis of intuitive arguments, since it is not a heuristic procedure. Through the unsupervised learning process of FCGD-B, we can obtain the cluster centers and fuzzy membership information for the given data set. In OFCAM, this algorithm is iterated for increasing number of clusters in a given interval, computing the value of validity measure[3] in each run. The validity measure, I_G computes the overall average compactness and separation of a fuzzy c partition obtained by FCGD-B. And this model use a validity strategy[3], which selects an optimal number of clusters presented in the data using the relative values and normalized value of the validity measure, I_G. The overall procedure of this model is summerized as followings:

step 1. Set the interval of the number of clusters, $I=[i_s, i_e]$.
Prepare input data set. $c=i_s-1$.

step 2. Find a fuzzy c-partition using FCGD-B.

step 3. Compute the value of cluster validity function, $I_G(c)$.

step 4. If $c=I_e+1$ then go to step 5.
Else increase c by 1 and go to step 2.

step 5. Select optimal number of clusters, c^*, by the validity strategy. And output c^* cluster centers and fuzzy membership information of c^* partitions.

Ⅲ. Construction of PCSDB using OFCAM

Considering the major components of the classification systems[6], we suggest a functional block diagram for the systems, as shown in Figure 1.

The data acquisition module and data preprocessing module, which are much more problem dependent than any other comonents, play a central role of the system. But, we focus on the learning module and

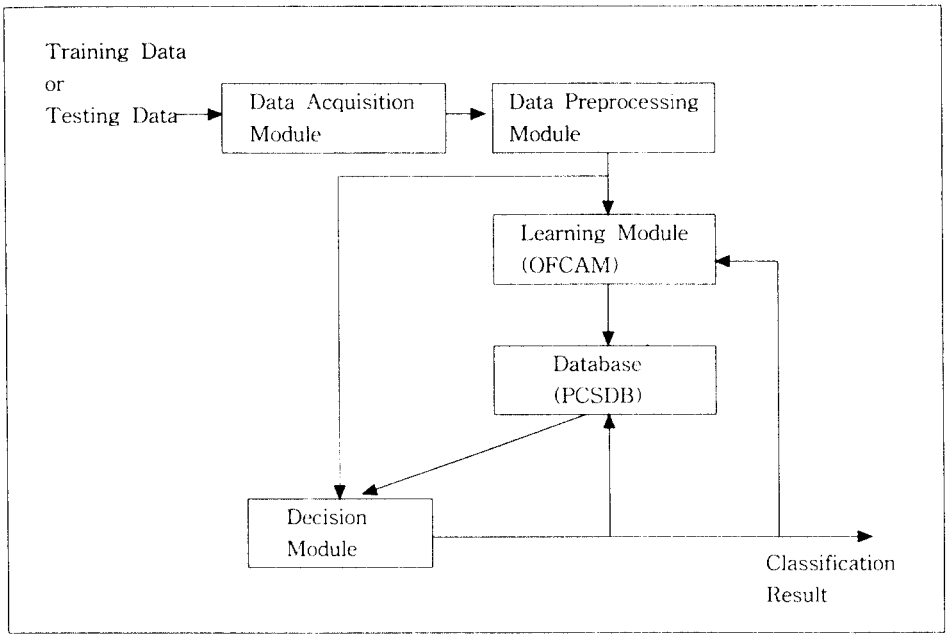

Fig. 1 Functional Block Diagram of a Pattern Classification System

decision module. In the conventional learning module, the labled samples must be given to provide *a priori* information for the classification. Moreover, the number of classes to be categorized must be known *a priori*. In our approach, we will use the model OFCAM, which was proposed in Section Ⅱ, as the learning module of the classification system. OFCAM automatically determines the optimal number of clusters, c^*, and c^* class information. And, the results from the OFCAM are stored in the database, called PCSDB, so as to be conviently used in the decision module and be systematically maintained in the classification system. Tables in PCSDB defined with the data definition language(DDL) statements of SQL are as follows:

```
CREATE TABLE DI (
DNAME          CHAR(10)      NOT NULL,
NCLASS         INTEGER       NOT NULL,
DIMENSION      INTEGER       NOT NULL,
NTP            INTEGER       NOT NULL,
PRIMARY KEY (DNAME));

CREATE TABLE DATAI (
DATA#          NUMERIC       NOT NULL,
DF1            NUMERIC       NOT NULL,
DF2            NUMERIC       NOT NULL,
...
DFP            NUMERIC       NOT NULL,
DATA_TYPE      CHAR          NOT NULL,
ACLASS         NUMERIC,
PRIMARY KEY (DATA#));

CREATE TABLE DATAI_CENTER (
CLASS#         NUMERIC       NOT NULL,
```

```
F1          NUMERIC     NOT NULL,
F2          NUMERIC     NOT NULL,
...
FP          NUMERIC     NOT NULL,
PRIMARY KEY (CLASS#));
```

* Data type NUMERIC is for numeric data; INTEGER, SMALLINT, FLOAT.

TABLE DI has four columns, DNAME which is the attribute for representing the name of any given data set, for example, DATAI, NCLASS for representing the number of classes of any given data set, which is determined by OFCAM, DIMENSION representing dimension of given data points, P in this example, NTP for representing the number of data points in the training data set, and having the indicated data types; column DNAME is the primary key. One tuple in TABLE DI indicates general information for any given data set. TABLE DATAI has P+3 columns, DATA# for the serial number for each data in DATAI, DF1, ···, DFP representing P dimensional input data item, D_TYPE which indicates whether the data representing the tuple is for training(L) or for testing(T), and also ACLASS which indicates a specific class which is assigned by the learning of OFCAM or decision process. The value of ACLAS can be NULL if it is not yet known. TABLE DATAI_CENTER has P+1 columns, CLASS# which is for the serial number for each class in the partition space, F1, ···, FP representing P dimension center point for each class and having indicated data types; column CLASS# is the primary key.

Once PCSDB has been constructed from the results of OFCAM, the system outputs the classification result for given test data using the information of PCSDB. The classification information of this model is stored into PCSDB and it is used as a feedback infomation to update the PCSDB by reflecting the testing data in the learning process. The decision process is as follows.

(1) Compute the membership degree which data point x is belonging to class i, $u_i(x)$. It is computed by the equation od Eq.(1).

$$u_i(x) = \frac{1}{\sum_{s=1}^{c} \left(\frac{\|x - v_i\|}{\|x - v_s\|}\right)^{\frac{2}{(m-1)}}} \quad (1)$$

, where v_i is the center point representing class i.

(2) Determine l with the largest value of $u_i(x)$, for $1 \leq i \leq c$ as a classification result of x

(3) Insert the test pattern x and classification result l as a new tuple into the TABLE DATAI.

Ⅳ. Application Example

To show that OFCAM can be effectively used in the proposed classification, we prepare 240 images, which were extracted from irregular images in nature[7]. They consist of eight classes each containing 30 images. Feature vectors for the images consist of 240 eleven dimensional vectors[7]. The results of running the OFCAM with this data set show that this data set consists of eight clusters as expected. And, the PCSDB constucted from the results has the TABLES defined in the previous section and each of them is shown as follows.

TABLE DATAI shows only for 15 data items, a part of given data set, called DATAI.

The performance of the pattern classification system designed by the above PRSDB is estimated by *leaving-one-out* method[8]. *Leaving-one-out* is an elegant and straightfoward technique for estimating classifier error rates. Because it is computationally expensive, it is often reserved for relatively small samples. For the sample size N, a classifier generated using N-1 patterns and tested on the remaining pattern. This is

TABLE DI

DNAME	NCLASS	DIMENSION	NTP
DATAI	8	11	240

TABLE DATAI

DATA#	DF1	DF2	DF3	DF4	DF5	DF6	DF7	DF8	DF9	DF10	DF11	D_TYPE	ACLASS
1	68	82	72	73	189	212	182	180	157	109	9	L	1
2	77	89	81	65	193	213	186	184	142	116	23	T	NULL
31	88	132	72	73	200	212	192	200	177	129	40	L	2
32	97	129	81	75	193	213	196	204	182	136	43	T	NULL
61	109	200	77	61	233	236	229	226	132	211	47	L	3
62	109	241	76	70	234	239	232	230	125	211	57	T	NULL
91	28	33	21	15	183	191	166	161	10	112	134	L	4
121	135	138	94	79	223	231	221	219	122	189	86	L	5
122	135	142	89	83	223	232	222	221	121	197	88	T	NULL
151	8	21	10	8	125	137	109	92	4	19	68	L	6
152	8	20	9	6	112	138	96	97	4	28	64	T	NULL
181	69	81	39	48	193	204	175	183	17	112	191	L	7
182	72	79	46	54	201	207	181	193	22	129	199	T	NULL
211	55	55	33	26	178	195	160	167	46	95	93	L	8
212	54	58	37	22	174	196	155	163	40	80	92	T	NULL

TABLE DATAI_CENTERS

CLASS#	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
1	73.53	89.47	74.88	58.95	194.75	214.34	188.73	185.15	151.64	121.21	26.75
2	87.77	126.47	73.32	62.11	195.29	214.47	189.47	186.91	178.30	124.56	49.56
3	101.72	223.40	91.62	81.46	236.67	239.98	233.08	231.98	141.57	216.24	55.03
4	22.95	31.24	19.12	15.64	181.96	194.81	164.29	164.35	9.76	109.22	142.75
5	136.72	142.21	88.94	78.98	224.15	232.11	222.18	221.19	119.19	191.12	88.11
6	9.39	24.02	12.44	8.77	131.93	148.39	113.21	104.10	6.65	26.31	67.72
7	69.21	75.07	41.89	48.21	202.84	192.98	179.06	180.39	18.25	111.48	208.06
8	52.79	68.01	41.22	25.95	189.24	204.85	174.91	178.61	54.96	91.08	89.99

repeated N times, each time designing a classifier by *leaving-one-out*. Each pattern is used for testing and each time nearly all the patterns are used to design a classifier. In this example, N is 30 for each of eight classes, 29 patterns of them are used for learning of the classification system and the remaining pattern is used for testing. Thus the total number of training data set is 232 and that of test data set is 8. This experiment was repeated 30 times. These experiments show a perfectly correct and clear classification for the testing data.

In our classification model, if we could get the optimal clustering of given data set from OFCAM, we expect that the error rates by *leaving-one-method* is approximately to 0%. So, we can note that the performance of the classification system depend on the clustering result of OFCAM. And, this model can manipulate together the classification problems of various applications by TABLES in PCSDB to be independently constructed.

V. Conclusions

We have presented the basic mechanism and an application example to show that an optimal cluster analysis model, OFCAM, is applicable to the database construction of classification system. In the design of the conventional classification systems, the labled samples must be given to provide *a priori* information for the classification. Moreover, the number

of classes to be categorized must be known *a priori*. However, in our approach, OFCAM has automatically determined the optimal number of clusters, c^* and c^* class information. And, the results from OFCAM was stored in the database, called PCSDB, so as to be conviently used in the decision module and be systematically maintained in the classification system. The classification model with PCSDB has estimated by *leaving-one-out* method. The results have shown a perfectly correct and clear classification for the testing data. And we have noted that the performance of the classification system depend on the clustering results of OFCAM.

References

1. R. Duda and P. Hartm Pattern Classification and Scene Analysis, Wiley: New York, 1973.
2. H. S. Rhee and K. W. Oh, "Unsupervised Learning Network Based on Gradient Descent Procedure of Fuzzy Objective Function", Proc. of IEEE International Conference on Neural Networks, pp. 1427-1432, Washington, DC., 1996.
3. H. S. Rhee and K. W. Oh, "A Validity Measure for Fuzzy Clustering and Its Use in Selecting Optimal Number of Clusters", Proc. of IEEE International Conference on Fuzzy Systems(FUZZ-IEEE'96), New Orleans, Sep. 1996.
4. J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithm, Plenum Press: New York, 1981.
5. T. Kohonen, Self-Organizing Maps, Springer: New York, 1995.
6. S. T. Bow, Pattern Recognition and Image Preprocessing, Marcel Dekker, Inc., 1992.
7. 강 명구, "표면 결함 분류를 위한 자동적인 특징평가 및 판단 알고리즘", 포항공과대학 석사학위 논문, 1993.
8. K. Fucunaga and R. R. Hayes, "Estimation of Classifier Performance", IEEE Trans. on PAMI, Vol. 11, pp. 1087-1101, 1989.

이 현 숙(Hyun-Sook Rhee)정회원

1989년 2월 : 서강대학교 전자계산학과(학사)

1991년 2월 : 포항공과대학교 대학원 전자계산학과(공학석사)

1991년 2월~1993년 3월 : 한국전자통신연구소 컴퓨터연구단 연구원

1997년 2월 : 서강대학교 전자계산학과(공학박사)

1997년 3월~현재 : 동양공업전문대학 전산경영기술공학부 조교수

※주관심분야 : 퍼지 신경망 모델, 소프트 컴퓨팅, 영상정보처리, 소프트웨어 개발 방법론