IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

Feature Dimension Reduction Using Euclidean Distance Oriented Similarity Based Rough Set Model

A.C. Mondal[1], S. Kolay[2*]

[1]Department of Computer Science, University of Burdwan, Burdwan, India
[2]Electrical & Automation, SMS India Pvt. Ltd., Kolkata, India

[*]*Corresponding Author: kolaysrikanta@gmail.com, Tel.: +91-98368-05404*

Available online at: www.ijcseonline.org

***Abstract*—** In machine learning, a very high dimensional data reduces the performance of a classifier. To overcome this, a relevant feature dimension reduction algorithm can be applied before applying any classification algorithm. Rough set theory [1] is a very good tool to reduce the feature dimension of an information system or decision system. However, if a decision system contains real-valued data, we cannot apply directly the rough set theory. Various extensions to rough set can be used to handle this kind of data. Among them, Fuzzy-Rough set theory [2], similarity based rough set model [3] are interesting. We propose an algorithm for dimension reduction using Euclidean distance oriented similarity based rough set model. To show the effectiveness of the algorithm, we take Grammatical Facial Expression Dataset from UCI Machine Learning Repository, created by Freitas et al. [4] and applied KNN classifier before and after feature dimension reduction.

***Keywords*—**Rough Set, Feature Dimension Reduction, Similarity Relation, KNN, Grammatical Facial Expression Recognition

I. INTRODUCTION

Rough set theory [1] was proposed by Zdzisław I. Pawlak in 1991. Rough set is a formal approximation of crisp set, described by lower approximation and upper approximation. Rough set theory can be used as a good tool for feature dimension reduction of an information system or decision system. However, we cannot apply rough set theory directly on an information system or decision system having real-valued data. To overcome this situation there are many extensions on rough set theory is available, e.g., fuzzy-rough set theory [2], similarity based rough set theory [3] etc. In this paper we discuss Euclidean distance oriented similarity based rough set model. It is very well known fact that performance of a classifier highly depends on the dimension of the given data set. Here we show the performance of KNN classifier before and after dimension reduction using Euclidean distance oriented similarity based rough set model. Rest of the paper is organized as follows, Section I contains the introduction, Section II contains the concept of rough set theory, Section III contains the concept of similarity based rough set model, Section IV describes the performance of KNN classifier on grammatical facial expression recognition before and after feature dimension reduction using Euclidean distance oriented similarity based rough set model, and section V contains the conclusion and future scope.

II. ROUGH SET THEORY

In this section, we recall the basic rough set theory [2,5] and its related terms. To define rough set first we need to define information system and decision system.

A. Information System

An *information system* can be defined as $I = (U, A)$, where U indicates finite, non-empty set of objects and A indicates finite, non-empty set of attributes, i.e. a: $U \rightarrow V_a$ for $a \in A$, where V_a is called the *value set* of a.

B. Decision System

In a *decision system* $A = C \cup D$, where C is the set of conditional attributes and D is the set of decision attributes.

C. Rough Set

Let, $P \subseteq A$, the equivalence relation can be defined as follows:

$IND(P) = \{(x,y) \in U^2 \mid \forall a \in P, a(x) = a(y)\}$ [1]

We can calculate the partition of U as follows:

$U/P = \otimes\{a \in P: U/IND(\{a\})\}$

where $A \otimes B = \{X \cap Y: \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\}$.

If $(x_1,x_2) \in IND(P)$, then x_1 and x_2 are indiscernible by the attribute subset P. The equivalence classes of P-indiscernibility relation can be denoted by $[x]_P$.

D. Lower Approximation

Let $X \subseteq U$, the P-lower approximation of a set can now be defined as:

$\underline{P}X = \{x \mid [x]_P \subseteq X\}$

E. Upper Approximation

P-upper approximation of a set can be defined as:

$\overline{P}X = \{x \mid [x]_P \cap X \neq \emptyset\}$

F. Rough Set

Rough set is defined by the pair $<\underline{P}X, \overline{P}X>$.

G. Positive Region

Let us assume that *P* and *Q* are the equivalence relations over *U*. The positive region *P* over *Q* can be defined as:

$POS_p(Q) = \cup_{X \in U/Q} \underline{P}X$

H. Dependency Factor

Let, $P,Q \subseteq A$, the dependency factor $\gamma_p(Q)$ can be defined as:

$\gamma_p(Q) = |Pos_p(Q)| / |U|$ where $|U|$ indicates the cardinality of set *U* and $|Pos_p(Q)|$ indicates the cardinality of $Pos_p(Q)$.

I. Dimension Reduction using Rough Set

Let us take an example of a small decision system [5] shown in Table 1.

Table 1. Decision System

	a	***b***	***c***	***d***	***q***
x_1	1	0	1	2	0
x_2	2	1	2	2	2
x_3	1	1	2	1	1
x_4	1	2	2	1	2
x_5	0	0	1	2	1
x_6	1	2	0	1	2

Here, $U=\{x_1, x_2, x_3, x_4, x_5, x_6\}$, $A=\{a, b, c, d, q\}$ where $C=\{a, b, c, d\}$ and $D=\{q\}$.

Considering $Q=\{q\}$ we get the partition of *U* as below:

$U/Q=\{X_1, X_2, X_3\}$ Where $X_1=\{x_1\}$, $X_2=\{x_2, x_4, x_6\}$, $X_3=\{x_3, x_5\}$.

Now, considering $P=\{a\}$, we get the equivalence classes as:

$U/P=\{\{x_1, x_3, x_4, x_6\}, \{x_2\}, \{x_5\}\}$.

Let us now find the lower approximations as:

$\underline{P}X_1=\{\emptyset\}$, $\underline{P}X_2=\{x_2\}$, $\underline{P}X_3=\{x_5\}$.

The positive region $POS_p(Q) = \cup_{X \in U/Q} \underline{P}X = \{x_2, x_5\}$.

Hence, $\{q\}$ depends on $\{a\}$ in a degree $\gamma_{\{a\}}(\{q\})$

where $\gamma_{\{a\}}(\{q\}) = |Pos_{\{a\}}(\{q\})| / |U|$

$= |\{x_2, x_5\}| / |\{x_1, x_2, x_3, x_4, x_5, x_6\}| = 2/6$.

In the similar fashion, we can calculate the dependencies for all possible subsets of *C* as:

$\gamma_{\{a\}}(\{q\})=2/6$, $\gamma_{\{b\}}(\{q\})=0$, $\gamma_{\{c\}}(\{q\})=1/6$,
$\gamma_{\{d\}}(\{q\})=0$, $\gamma_{\{a,b\}}(\{q\})=1$, $\gamma_{\{a,c\}}(\{q\})=4/6$,
$\gamma_{\{a,d\}}(\{q\})=3/6$, $\gamma_{\{b,c\}}(\{q\})=2/6$, $\gamma_{\{b,d\}}(\{q\})=4/6$,
$\gamma_{\{c,d\}}(\{q\})=2/6$, $\gamma_{\{a,b,c\}}(\{q\})=1$, $\gamma_{\{a,b,d\}}(\{q\})=1$,
$\gamma_{\{a,c,d\}}(\{q\})=4/6$, $\gamma_{\{b,c,d\}}(\{q\})=4/6$, $\gamma_{\{a,b,c,d\}}(\{q\})=1$.

Here $\{a,b\}$ is the minimal subset of the conditional attributes for which dependency factor is equal to 1.

Hence, $\{a,b\}$ is one reduct.

J. Dimension Reduction using Rough Set when Attribute Values Contains Real-Valued Data

Let us now modify the above example a little as shown in Table 2.

Table 2. Modified Decision System

	a	***b***	***c***	***d***	***q***
x_1	0.99	0	1	2	0
x_2	2	1	2	2	2
x_3	1.01	1	2	1	1
x_4	1	2	2	1	2
x_5	0	0	1	2	1
x_6	1	2	0	1	2

If we now calculate the reduct using rough set theory directly, we find $\{a\}$ as one reduct.

As per our intuition, it is clear that $\{a\}$ should not be a reduct. So, how to overcome this situation? One of the solutions is similarity based rough set model.

III. SIMILARITY BASED ROUGH SET MODEL

In similarity based rough set model [3] instead of an indiscernibility relation a similarity relation is used.

A. Euclidean Distance Oriented Similarity Based Rough Set Model

Here, similarity between two objects is measured through the Euclidean distance. To use Euclidean distance measure we need to normalize the data set.

Let $B \subseteq A$, to construct global similarity relation we can define $xSIM_By$ *iff* $ED_B(x,y) \geq t$, where $ED_B(x,y)$ indicates the Euclidean distance between *x* and *y* based on the attributes *B* and *t* indicates a predefined threshold.

B. Algorithm to Find a Reduct Using Euclidean Distance Oriented Similarity Based Rough Set Model

$C \leftarrow$ Conditional attribute set
$D \leftarrow$ Decisional attribute set
Step-1: Normalize the decision system
Step-2: Set the value of t
Step-3: $R \leftarrow \{\}$
Step-4: Do
Step-5: $T \leftarrow R$
Step-6: $\forall x \in (C-R)$
Step-7: if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
Step-8: $T \leftarrow R \cup \{x\}$
Step-9: $R \leftarrow T$
Step-10: Until $\gamma_R(D) \geq \gamma_C(D)$
Step-11: Return *R*

IV. GRAMMATICAL FACIAL EXPRESSION RECOGNITION

Here, we take Grammatical Facial Expression Dataset from UCI Machine Learning Repository, created by Freitas et al. [4]. The conditional attributes containing one hundred coordinates (x, y, z) of points from eyes, nose, eyebrows, face contour and iris as shown in Figure 1 below.

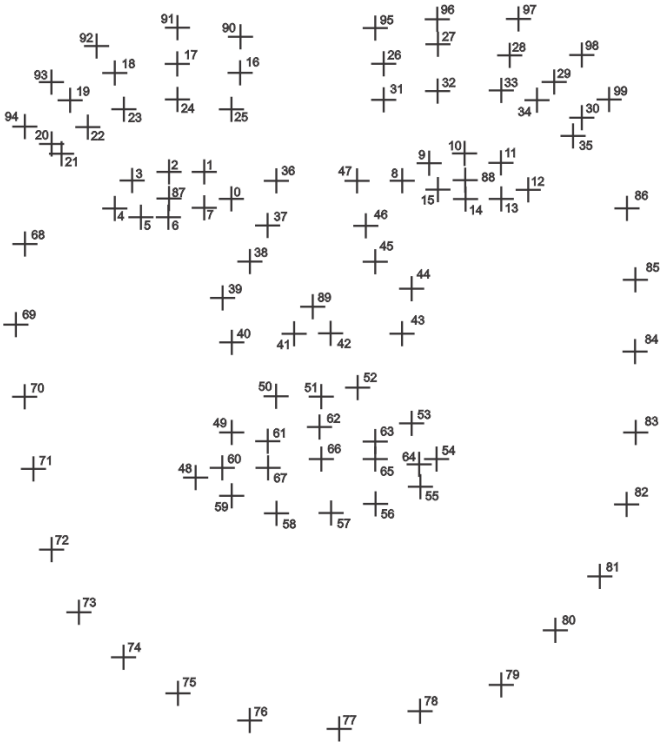

Figure 1. Attribute point locations on user face

Based on the conditional attributes, the decisional attributes are as follows:

1. Affirmative
2. Conditional
3. Doubt Question
4. Emphasis
5. Negative
6. Relative
7. Topics
8. WH Question
9. YN Question

We took only the positive samples for our classification. We consider 80% data of each class as training data and 20% data of each class as test data. Now we use KNN classifier for classification purpose. The performance of the classifier (taking K=3) is shown in Table 3 below:

Table 3. KNN performance on complete dimension of data

Decision Class	Total Number of Samples	Number of Successful Classification	Success Rate (%)
Affirmative	94	81	86.17
Conditional	114	92	80.7
Doubt Question	127	116	91.33
Emphasis	86	71	82.55
Negative	124	98	79.03
Relative	119	90	75.63
Topics	83	56	67.46
WH Question	116	113	97.41
YN Question	124	106	85.48
Total:	987	823	83.38

Now, if we apply Euclidean distance oriented similarity based rough set model to find a reduct using our algorithm, we find the attribute subset {9, 25, 82} as reduct.

Based on this reduced dimension of data if we apply KNN classifier (taking K=3), we get the result as shown in Table 4 below:

Table 3. KNN performance on reduced dimension of data

Decision Class	Total Number of Samples	Number of Successful Classification	Success Rate (%)
Affirmative	94	83	88.3
Conditional	114	93	81.58
Doubt Question	127	122	96.06
Emphasis	86	80	93.02
Negative	124	115	92.74
Relative	119	103	86.55
Topics	83	71	85.54
WH Question	116	115	99.14
YN Question	124	118	95.16
Total:	987	900	91.18

So, here we can clearly see the benefit of our dimension reduction algorithm for KNN classifier.

V. CONCLUSION

This paper focused on standard rough set model and similarity based rough set model. We proposed an algorithm to find a reduct using Euclidean distance oriented similarity based rough set model. Finally we showed its effectiveness on KNN classifier. The reason for selecting KNN classifier is, both Euclidean distance based similarity based rough set model and KNN classifier use the distance measure between the objects. In future we shall make a comparative study by applying the result of the dimension reduction on other classifiers.

REFERENCES

[1] Z. Pawlak, "*Rough Sets. Theoretical Aspects of Reasoning about Data*", Kluwer Academic Publishers, 1991.

[2] R. Jensen, Q. Shen, "*Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches*", IEEE Transactions on Knowledge and Data Engineering, Vol.16, Issue.12, pp.1457-1471, 2004.

[3] J. Stepaniuk, "*Similarity Based Rough Sets and Learning*", Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, Tokyo, Japan, pp. 18-22, 1996.

[4] F. A. Freitas, S. M.. Peres, C. A. M. Lima, F. V. Barbosa, "*Grammatical Facial Expressions Recognition with Machine Learning*", 27th Florida Artificial Intelligence Research Society Conference (FLAIRS), Palo Alto, pp. 180-185, 2014.

[5] K. S. Ray, S. Kolay "*Application of Approximate Equality for Reduction of Feature Vector Dimension*", Journal of Pattern Recognition Research, Vol.11, Issue. 1, pp.26-40, 2016

Authors Profile

Dr. Abhoy Chand Mondal is currently Professor of the Department of Computer Science, The University of Burdwan, W.B., India. He received his B.Sc.(Mathematics Hons.) from The University of Burdwan in 1987, M.Sc. (Math) and M.C.A. from Jadavpur University, in 1989, 1992 respectively. He received his Ph.D. from Burdwan University in 2004. He has 1 year industry experience and 21 years of teaching and research experience. No. of papers more than 70 and no of journal is more than 30. His research interest includes fuzzy logic, soft computing, document processing, natural language processing, big data analitics etc.

Mr. Srikanta Kolay is currently working as Deputy General Manager in SMS India Pvt. Ltd. He received his B.Sc. (H) in Computer Science and and Master of Computer Application (MCA) degree 2003 and 2006 respectively from Vidyasagar University, Midnapore, India. He has two research papres in international journals. His research interest includes rough set, fuzzy logic, pattern recognition, data mining etc.

