IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

PERFORMANCE ENHANCEMENT OF DISTANCED BASED ALGORITHMS FOR CLASSIFICATION PROCESS

Mrs. A. Sumathi[*1] & Dr. N. Sengottaiyan[2]
[*1]Assistant Professor, Department of Computer Science, Navarasam Arts & Science College for women, Erode, India
and
Part–Time Ph.D (Category – B), Research and Development Centre, Bharathiar University, Coimbatore, India.
[2]Director, Hindusthan College of Engineering & Technology, Coimbatore, India

DOI: 10.5281/zenodo.839133

ABSTRACT

Nowadays there is vast amount of data being collected and stored in databases and without automatic methods for extracting this information it is practically impossible to mine for them. In Data Mining, the Classification processes perchance the most recognizable and most popular concepts. Actually Classification maps data into predefined groups or classes. Classification normally uses prediction rules to express knowledge. This Prediction rules are expressed in the form of IF - THEN rules. Sometimes this classification referred to as supervised learning by reason of the classes is determined before examining the data. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. Classification problems handled by using some known type of classification algorithms such that Statistical-Based Algorithms, Distance Based Algorithms etc,. All approaches to performing classification assume some knowledge of the data. This paper will focus on Distance Based Algorithms in classification. In Distance Based Algorithms each item is mapped to the same class may be thought of as more similar to the other items in that class than items found in other classes. Therefore, similarity or distance measures may be used to identify the "alikeness" of different items in the database. Using a similarity measure for classification where the classes are predefined is somewhat simpler than using a similarity measure for clustering where the classes are not known in advance. So the classification problem then becomes one of determining similarity not among all tuples in the database but between each tuple and the query

KEYWORDS: Classification, Similarity, Tuples, Attributes.

I. INTRODUCTION

The idea of similarity measure can be abstracted and applied to more general classification problems. The difficulty lies in how the similarity measures are defined and applied to the items in the database. Since most similarity measures assume numeric values, they might be difficult to use for more general or abstract data types. A mapping from the attribute domain to a subset of integers may be used.

The Distance Based Algorithms consists of 2 approaches:

i) Simple Approach
ii) K Nearest Neighbors(KNN)

Simple Approach

Problem Description

This approach uses the Information Retrieval (IR) approach to assigning each tuple to the class to which it is most similar. Here, the tuple is represented by t_i in the database. The database defines a set of vectors like <ti1,ti2,......,t_{ik}>.The vector contains only numeric values.

Definition

Given a database D = $\{t_1, t_2, \ldots., t_n\}$ of tuples where each tuple $t_i = \{t_{i1}, t_{i2}, \ldots., t_{ik}\}$ contains numeric values and a set of classes C = $\{C_1, C_2, \ldots., C_m\}$ where each class $C_j = <C_{j1}, C_{j2}, \ldots. C_{jk}>$ has numeric values, the classification problem is to assign each t_i to the class C_j such that $(t_i, C_j) \geq \text{sim}(t_i, C_l)$ ∀ Cl ∈ C where $C_l \neq C_j$.

Measurement

To calculate similarity measures, each class representative must be determined.

Consider three classes that Class A, Class B and Class C. We can calculate the center of each region to determine a representative for each class. A simple classification technique is placing each item in the class where it is most similar or closest to the center of that class.

Algorithm

Input:

```
c1...., cm          //centers for each class
t           //Input tuple to classify
```

Output:

```
c           //class to which t is assigned
Simple distance-based algorithm.
dist = ∞;
for i := 1 to m do
if dis (ci, t) < dist, then
c = i;
dist = dist(ci,t);
```

Algorithm Explanation

- This algorithm illustrates approach of distance-based to identify center or centroid value c_i.
- Since each tuple must be compared to the center for a class and there are a fixed number of classes.
- The complexity of one tuple is **O(n)**.

Sample Results

- This figure displays the use of this simple approach to perform classification.
- The three large dark circles are the class representatives for the three classes.
- The dashed lines represent distance from each item to the closet center.
- The tree classes are Class A, Class B and Class C.

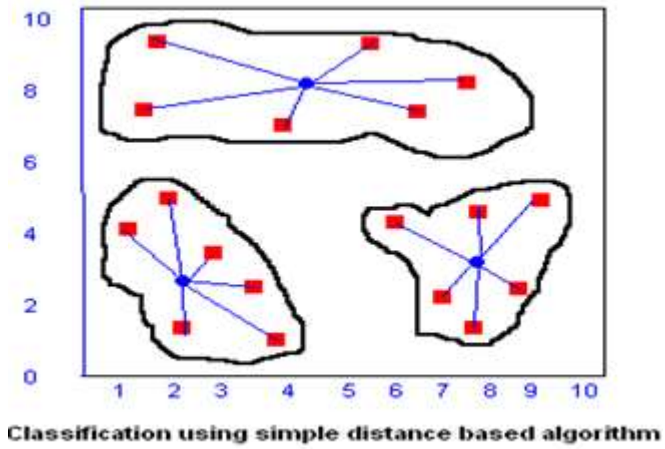

Figure 1. Classification using simple Distance Based Algorithms

K Nearest Neighbors(KNN)

Definition

Given a query point q for which we want to know its class l, and a training set $X = \{\{x_1, l_1\}...\{x_n\}\}$, where x_j is the j-th element and l_j is its class label, the k-nearest neighbors will find a subset $Y = \{\{y_1, l_1\}...\{y_k\}\}$ such that $Y \in X$ and $\Sigma^{k}_{1} d(q,y_k)$ is minimal. Y contains the k points in X which are closest to the query point q. Then, the class label of q is $l = f(\{l_1...l_k\})$.

It is a one of the common classification scheme based on the use of distance measures. It assumes that the entire training set includes desired classification for each item not only the data in the set.

Whenever the classification is made for a new item, its distance to each item in the training set must be determined. After that, the K closet entries in the training set are considered further.

The new item is then placed in the class that contains the most items from this set K closet items. *k*-NN is a special case of a variable-bandwidth, kernel density "balloon" estimator with a uniform kernel.

Algorithm

Input:

```
T          //Training data
K           //Number of neighbours
t          //Input tuple to classify
```

Output:

```
c          //Class to which t is assigned
```

KNN algorithm:

```
//Algorithm to classify tuple using KNN
N = ∅;
//Find set of neighbors, N, for t
for each d Є T do
if | N |≤ K, then
N = N ∪ {d};
else
if Ǝ u Є N such that
          sim(t, u) ≤ sim(t, d), then
begin
N = N - {u};
N = N ∪ {d};
end
//Find class for classification
c = class to which the most u ∈ N are classified;
```

Algorithm Explanation

- In this algorithm we use T to represent the training data. Since each tuple to be classified must be compared to each element in the training data.
- If there are q elements in the training set, this is O(q).
- Given n elements to be classified, this becomes an O(nq) problem.
- Given that the training data are constant size, it can be viewed as as O(n) problem.

Sample Results

- This figure illustrates the process used by KNN.
- The training set points are shown here.
- Then K value is 3. So three closet items in the training set are shown here.
- The tuple t placed in the class that most of these members.

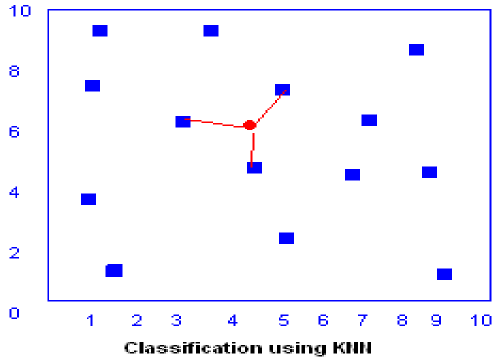

Figure 2: Classification using KNN

In this paper we focus on the *k*-nearest-neighbor search method, which, until recently was not considered for small molecule classification. The few recent applications of *k*-nn to compound classification focus on selecting the most relevant set of chemical descriptors which are then compared under standard Minkowski distance L_p. Here we show how to computationally design the optimal *weighted* Minkowski distance wL_p for maximizing the discrimination between active and inactive compounds wrt bioactivities of interest.

We then show how to construct pruning based *k*-nn search data structures for any wL_p distance that minimizes similarity search time.

The accuracy achieved by our classifier is better than the alternative LDA and MLR approaches and is comparable to the ANN methods. In terms of running time, our classifier is considerably faster than the ANN approach especially when large data sets are used. Furthermore, our classifier quantifies the level of bioactivity rather than returning a binary decision and thus is more informative than the ANN approach.

k-NN has some strong consistency results. As the amount of data approaches infinity, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data)

In the same way we assume each class C_j is defined by a tuple <C_{j1}, C_{j2} ,.... C_{jk}>. The class also includes numeric values.

II. RELATED WORKS

There are a large number of technologies which have been proposed for efficiently Distance based algorithm on classification.

However, no proposed work concerns to selecting an appropriate scale. Our work is similar in spirit to two similarity based algorithm.

Our solution is different with the solution in on 2 aspects. One is algorithm selection and another one is decomposition of data.

III. CONCLUSION

In data mining, classification was frequently formed by simply applying knowledge of the data. The classification problem consists of different approaches to find prediction. The simple and easiest method is Distance-Based algorithms. This algorithm consists of two approaches like Simple Approach and K Nearest Neighbors(KNN) to easily predict the future outcomes based on the given input values.

Finally, these papers conclude that the distance based algorithms are more helpful to find out the prediction for classification problems. We believe that identifying DB-outliers is an important and useful data mining activity. In this paper, we proposed and analyzed two algorithms for finding classification prediction. The proposed classification method based on cluster and distance to find prediction. Also using distance based algorithms reduces the size of the problem and provides a better result. In our future work we are going to continue validation of the proposed method.

IV. FUTURE WORK

We intend to extend this work to two directions. First, applying this approach to a real time series dataset. The target of mining this data set includes time series classification. Second, combining these work with distance-based clustering algorithm. In fact, the main idea of this work is selecting appropriate features of an algorithm for better Distance comparison

V. REFERENCES

[1] Margaret H.Dunbam – “Data Mining Introductory and Advanced Topics” Pearson Education – 2003.
[2] Jiawei Han & Micheline Kamber – “Data Mining Concepts and Techniques” 2001, Academic press.
[3] “Data Mining and Knowledge Discovery with Evolutionary Algorithms”, A.A. Freitas, Springer-Verlag,2002.
[4] Tan P.-N., Steinbach M. and Kumar V., “Introduction to Data Mining” , Addison Wesley, 2006.
[5] Hand D., Mannila H. and Smyth P., “Principles of Data Mining”, MIT Press, 2001.
[6] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician 46 (3): 175–185.
[7] Hall P, Park BU, Samworth RJ (2008). "Choice of neighbor order in nearest-neighbor classification". Annals of Statistics 36 (5): 2135–2152.
[8] Cover TM, Hart PE (1967). "Nearest neighbor pattern classification". IEEE Transactions on Information Theory 13 (1): 21–27.
[9] T.M.Cover, "Estimation by the nearest neighbor rule", IEEE Trans. Inf. Theory IT-14 (1) (1968)50–55.
[10] T.M. Cover, P.E. Hart, "Nearest neighbor pattern classification", IEEE Trans. Inf. Theory IT - 13 (1) (1967) 21 – 27.
[11] J.M. Keller, M.R. Gray, J.A. Givens Jr., "A fuzzy K - nearest neighbor algorithm", IEEE Trans. Syst. Man Cybern. SMC - 15 (4) (1985) 580 – 585.
[12] M. Sarkar, "Fuzzy - rough nearest neighbor algorithms in classification", Fuzzy Sets and Systems 158 (2007) 2134–2152..

CITE AN ARTICLE

Sumathi, A., and N. Sengottaiyan. "PERFORMANCE ENHANCEMENT OF DISTANCED BASED ALGORITHMS FOR CLASSIFICATION PROCESS." ***INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*** **6.8 (2017): 104-08. Web. 5 Aug. 2017.**