Towards the new Similarity Measures in Application of Machine Learning Techniques on Agriculture Dataset

Bhagirath Parshuram Prajapati
Department of Computer Engineering
A. D. Patel Institute of Technology
New, V. V. Nagar

Dhaval R. Kathiriya
Agricultural Information Technology
College of Agricultural Information Technology
Anand, India

ABSTRACT
k-Nearest Neighbor is a simple and effective classification method. The primary idea of this method is to calculate the distance from a query point to all of classified data points and make choice of a class which occurs maximum time in k closest neighbors. The Euclidean distance and cosine similarity the common choice for similarity metric among all the similarity measures. Apart from Euclidean and Cosine there are various similarity measures available and being used to calculate similarity in n-dimension vector space model for classification. Similarity calculation is complex operation and computationally need high time if vector dimension increases. Hence this paper explores the usefulness of nine different similarity measures in kNN and presents their experimental results on agriculture dataset. We also compared the time required to finish the classification task and concluded that I-divergence is taking minimum time compared to these algorithms.

General Terms
k-Nearest Neighbor, Similarity measure

Keywords
Euclidian, Manhattan, Minkowasky, Canberra, Chebychev, Cosine, Correlation, Chi-square, I-divergence

1. INTRODUCTION
In this paper we have explored a machine learning approach to classify agriculture data into predefined categories. This classification method is supported by the intelligent machine learning and data mining algorithms rather than human annonator's hard work . The classification activity is carried out using a K Nearest Neighbor (kNN) classification method. K-Nearest Neighbor (kNN) is a simple, effective and nonparametric method, which is a machine learning algorithm without building any generative models. kNN is a classification method based on the knowledge learned from the existing classified records into a set of categories. In this method, top k nearest neighbors data points to the test data point are found and then the class of the test data point is assigned to a class which is most present in the set of nearest neighbor. KNN has been used in many areas for example, text classification, pattern recognition and biometric etc.

One of the major challenge in classification is selection of distance measure to calculate the distance between data points. This paper presents nine simple and efficient ways to compute similarity measures. We achieve state of the art performance on agriculture datasets.

This paper is organized as follows. Section 2 presents the related work. Section 3 describes the nine similarity measures. The experiment and evaluation is illustrated in Section 4. The outcomes of analysis that we carried out based on evaluation is presented in Section 5. Section 6 concludes this paper and provides some avenues for future directions.

2. RELATED WORK
Many researches have provided a survey on comparison of different similarity measures for different application. One of the major paper in this area is Similar observations were also reported in the Zobel, J. and A. Moffat's paper, [10] on text mining. In one work, there is consideration of time taken to classify [11]. To the best of our knowledge, we have not refer a single paper applied on agriculture dataset using different similarity measures.

2.1 kNN Classifier
The simple kNN classifier algorithm is presented below:

Step 1: Given a dataset $X = \{(x_i,c_i)|$ where $i = 1,2,..,n\}$ where xi denotes a data point consist of feature vector denoted by xi $= (x_i^1,x_i^2,\ldots,x_i^m)$ where m is total number of features and x_i^k denotes a feature value where k = 1, 2, ..., m. c_i is the corresponding class of the i^{th} data point. And category ci belongs to category set $C = \{c_1, c_2, \ldots, c_t\}$ where t is total number of categories.

Step 2: Divide dataset into the training set $T = \{(x_i,c_i)|$ where $i = 1,2,..,n_1\}$ where n1 is number of training data points and test set $S = \{(x_i,c_i)|$ where $i = 1,2,..,n_2\}$ where n2 is number of test data points.

Step 3: Estimate initial value of k in kNN

Step 4: Calculate the distance between test data points and training data points using one of the distance measure presented in Section 3 and sort the distances in ascending or descending order according to the distance measure. Select the k data points with k relative distances as k nearest neighbours;

Step 5: Find the class which occurs maximum number of times in top k nearest neighbours and assign the class label to the test data point.

Step 6: Construct a confusion matrix and calculate accuracy from the confusion matrix.

3. SIMILARITY FUNCTION
In instance based learning algorithm like k-NN the distance function is used to decide which neighbors are closet to an input vector. Here in this research work, we have employed nine distance measures to evaluate the k-NN accuracy and the time it takes to complete classification task.

A metric or distance function is a function *D(X,Y)* that defines the distance between elements of a set as a non-negative real number. Distance zero means both elements are equal. To measure closeness of two element it is done by distance

function, here elements do not have to be numbers but can also be vectors, matrices or arbitrary objects

3.1 Euclidian similarity measure

It is straight line distance between two points in Euclidian space, it is derived from Pythagoras metric [1],

$$D(X,Y) = \sqrt{\sum_{i=1}^{m} (xi - yi)^2} \quad (1)$$

3.2 Manhattan similarity measure

It is distance between points of city road grid hence it is also known as city block distance [2], it is used to examine absolute difference between two points of object. The Manhattan distance is the simple sum of the horizontal and vertical components and hence it is always greater than or equal to zero.

$$D(X,Y) = \sum_{i=1}^{m} |xi - yi| \quad (2)$$

3.3 Minkowasky similarity measure

This distance measure is generalization of Euclidian and Manhattan similarity measure.

$$D(X,Y) = \left(\sum_{i=1}^{m} |xi - yi|^r \right)^{1/r} \quad (3)$$

3.4 Canberra similarity measure

Is it weighed version of Manhattan similarity measure, it is used for data scattered around the origin [3].

$$D(X,Y) = \sum_{i=1}^{m} \frac{|xi - yi|}{|xi + yi|} \quad (4)$$

3.5 Chebychev similarity measure

It is defined on a vector space where the distance between to vectors is the greatest of their differences along any coordinate dimension.

$$D(X,Y) = \max_{i=1}^{m} |xi - yi| \quad (5)$$

3.6 Cosine similarity measure

It measures the cosine angle between two non-zero vectors of an inner product space. If two vectors are having same orientation means cosine similarity of 1, and if orientation is at 90 means similarity 0. Diametrically opposite vectors have a similarity 1 [2].

$$S_{\cos} = \frac{\sum_{i=1}^{m} xiyi}{\sqrt{\sum_{i=1}^{m} xi^2} \sqrt{\sum_{i=1}^{m} yi^2}} \quad (6)$$

3.7 Correlation similarity measure

To quantify correlation between similarity measures, a correlation similarity is used. Strength and direction between two distance measures are indicated by it. If the value gets close to 1, it represents a good fit, i.e., two distance measures are semantically similar [4]. Correlation coefficient approaches zero when the fit gets worse. When either two distance or two similarity measures are compared, the correlation coefficient is a positive value.

$$D(X,Y) = \frac{\sum_{i=1}^{m} (xi - \mu i)(yi - \mu i)}{\sqrt{\sum_{i=1}^{m} (xi - \mu i)^2 \sum_{i=1}^{m} (yi - \mu i)^2}} \quad (7)$$

3.8 Chi-square similarity measure

The Chi-square distance is calculated on relative counts, and it is standardizes by the mean and not by the variance [5].

$$D(X,Y) = \sum_{i=1}^{m} \frac{1}{sum_i} \left(\frac{xi}{size_x} - \frac{yi}{size_y} \right)^2 \quad (8)$$

3.9 I-divergence similarity measure

It is usually used in positive, linear inverse problem, and it is a kind of distance metric showing the difference between measured value and true value [6].

$$D(X,Y) = \sum_{i=1}^{m} yi^i \log \frac{yi^i}{xi^i} - yi^i + xi^i \quad (9)$$

4. EXPERIMENTAL EVALUATION

4.1 Dataset

We have carried out our experiments on agriculture two data sets. One is Soil Health Card Dataset of Vadodara district. The dataset contains total 5183 records of soil type across Vadodara district. Each record has an entry of district id, taluka id, village id, soil id and quantity of Iron, Magazine, Copper and Zinc present in the soil. In addition to above data fields, one more field describing whether the soil has deficiency or not is added in the data set. The values of new field are "yes" or "no". The data set is divided into 50-50 % split of training and test set for experimental setup. Hence there are 2592 training records and 2591 test records on which task of classification was performed. The measure of effectiveness in our experiment is accuracy since the required information is whether a soil contains deficiency or not. Nine similarity measures were employed. The dataset is normalized into range from 0 to 1. A sample set of Soil Health Card Dataset of Vadodara district is present in Table1.

Table I. A sample set of Soil Health Card Dataset of Vadodara district for two class

VILLAGE_ID	SOIL_ID	SHC_IRON	SHC_MANGANESE	SHC_CU	SHC_ZINC	Deficiency
V13155	S10	12.68	13.32	0.72	1.42	No
V13136	S10	14.86	20.08	0.68	1.88	No
V12633	S10	3.6	9.2	0.18	0.68	Yes
V12411	S02	8.9	3.4	0.34	0.2	Yes
V12988	S13	8.4	9.84	0.38	0.38	Yes

V13162	S10	10.24	13.08	0.62	1.32	No
V12816	S02	15.26	12.26	0.54	1.44	No
V13155	S10	12.82	18.16	0.78	1.12	No

TABLE II. A sample set of Soil Health Card Dataset of Vadodara district for five class

SHC_Fe	SHC_Mn	SHC_Cu	SHC_Zn	SHC_S	SHC_Mg	Label
9.45	9.52	0.4	0.38	21	12.4	L3
9.87	6.7	0.34	0.3	35	12.4	L3
7.54	8	0.62	0.3	13	10.2	L9
10.04	7.16	0.74	0.46	6.3	1.6	L40
5.24	10.22	0.66	0.32	7.65	0.8	L24
9.45	8.26	0.74	0.68	14	12.4	L9
9.24	9.36	0.3	0.9	24	12.8	L3

Table II indicates multiple deficiencies labeled by L3, L9, L24, and L40.

This second database we applied our algorithms is Vadodara district data which includes values of Iron (Fe), Manganese(Mn), Zinc(Zn), sulphur(S), Magazine(Mg). The total number of records are 1000, among them we have labelled some dataset as deficiency of combination of different chemical attributes. We have labelled them as L3, L9, L12 and L24 according to their combination of deficiency in a particular chemical atoms. There are 200 records for each class in the dataset. A sample set of Soil Health Card Dataset of Vadodara district is present in Table2.

4.2 Results

This section reports results that we have carried out using nine similarity measures in kNN classifier on two data sets mentioned in the data section. Table 3 reports the comparison of accuracy and time taken to complete the given task by k-NN using nine different similarity measures on two class agriculture dataset. The effectiveness of different similarity measures are presented in terms of accuracy and time required to complete the task. The Figure 1 shows comparison of accuracy achieved by kNN using nine different similarity measures using 21 as a value of k on two class agriculture dataset. Table 4 presents the comparison of accuracy achieved by k-NN on Vadodara Five class dataset. The time to complete the classification task is also presented in Table 4. Figure 2 shows comparison of accuracy achieved by kNN on five class dataset where 21 is set as a value of k. The next section gives insights of the results present in this section.

Table III. Comparison of results (in terms of accuracy and time) of kNN using nine different similarity measures applied on two class Vadodara agriculture data set.

No	Similarity measure	Accuracy (%)	Time (Mili Second)
1	Euclidian	92	34534
2	Manhattan	80	32345
3	Minkowasky	75	33871
4	Canberra	70	32571
5	Chebychev	67	34522
6	Cosine	90	33595
7	Correlation	93	36664
8	Chi-square	85	33454
9	I-divergence	94	29550

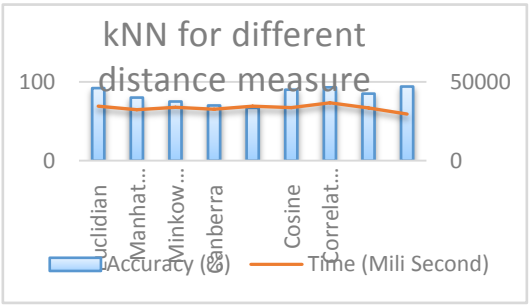

Fig. 1. The bar graph for Accuracy values for kNN using nine different similarity measures applied on two class agriculture dataset.

Table. IV Comparison of results (in terms of accuracy and time) of kNN using nine different similarity measures applied on five class Vadodara agriculture data set.

No	Similarity measure	Accuracy (%)	Time (Mili Second)
1	Euclidian	84	3730
2	Manhattan	73	3511
3	Minkowasky	77	3664
4	Canberra	64	3534
5	Chebychev	55	3729
6	Cosine	85	3636
7	Correlation	90	3943
8	Chi-square	80	3622
9	I-divergence	92	3232

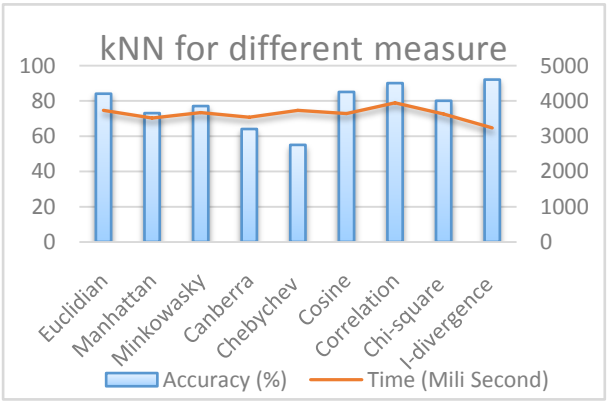

Fig. 2. The bar graph for Accuracy values for kNN using nine different similarity measures applied on five class agriculture dataset

5. ANALYSIS

In our experimental evaluation, we have observed that the performance comparison of all the nine different similarity measures (Euclidian, Manhattan, Minkowasky, Canberra, Chebychev, Cosine, Correlation, Chi-square, I-divergence) provides different effectiveness in terms of accuracy and time on both datasets. The I-divergence distance measure is the best similarity measure in terms of time and accuracy on both datasets. The next best performance reported here is achieved by Cosine, Correlation and Euclidian in terms of accuracy. But there is no significance difference in the all four distance measures.

As expected, I-divergence, Correlation and Cosine all had fairly good accuracy but I-divergence is taking minimum time compared to these algorithms. The reason behind the slowest time of I-divergence is that it we first store all required calculations prior to calculate final value of I-divergence for a particular test set. In doing so, we have reduced the time required to calculate various parameters for the desired output. It is really important to know how these various distances measures behave when applied on heterogeneous dataset.

There is an important research setting in determine value of k for k nearest neighbor The best choice of k depends upon the data, and can be chosen based on knowledge or experience about the classification task. A good value of k can be selected by various heuristic techniques such as cross-validation. In our experimental evaluation, we have observed that the performance comparison of Euclidean distance and cosine similarity has no significance difference. Similar observations were also reported in the [10], 1998 on comparison of similarity measures

6. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have explored nine different similarity measures in classification task. We have showed that the best choice for similarity measure is I-divergence since its effectiveness in term of time and accuracy better than other eight distance measures. The best choice of k depends upon the data, and can be chosen based on knowledge or experience about the classification task. A good value of k can be selected by various heuristic techniques such as cross-validation. The distance measures including Cosine, Correlation and Euclidian also performs better than remaining five distance measures.

The future directions that can be applied to agriculture field from aspect of machine learning and data mining is as follows. Association Mining useful to find important patterns and relation can be applied to agriculture dataset. A search engine for agriculture can be developed so a farmer can find useful information for his needs. For example, what type of crop should be grown in particular soil type given a particular session. We can do clustering in order to find new relationships among the data points.

7. REFERENCES

[1] Deza E. and Deza M.M., Dictionary of Distances, Elsevier, 2006

[2] Zezula P., Amato G., Dohnal V., and Batko M., Similarity Search The Metric Space Approach, Springer, 2006

[3] B.S.Charulatha, et. El, A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms, IJETTCS, National Conference on Architecture, Software systems and Green computing-2013(NCASG2013)

[4] Gavin D.G., Oswald W.W., Wahl, E.R., and Williams J.W., A statistical approach to evaluating distance metrics and analog assignments for pollen records, Quaternary Research 60, pp 356–367, 2003

[5] O. Ibrahimov, et. El, The performance analysis of a Chi-square similarity measure for topic related clustering of noisy transcripts, Pattern Recognition, 2002. Proceedings. 16th International Conference.

[6] Xiangyan Meng, et. El , A Novel K-Nearest Neighbor Algorithm Based on I-Divergence with application to Soil Moisture Estimation in Maize Field.

[7] R. Chang, Z. Pei, C. Zhang, "A Modified Editing k-nearest Neighbor Rule", Journal of Computers, vol.6, pp.1493-1500, 2011.

[8] J. Gou, T. Xiong, Y. Kuang, "A Novel Weighted Voting for K-Nearest Neighbor Rule", Journal of Computers, vol.6, pp.833-840, 2011.

[9] T.M. Cover, P. E. Hart, "Nearest Neighbor Pattern Classification", IEEE. Transactions on Information Theory, vol.13, no.1, pp.21-27, 1967.

[10] Zobel, J. and A. Moffat, Exploring the Similarity Space. In ACM SIGIR Forum. 1998.

[11] B. Prajapati, D. Kathiriya, "Evaluation of Effectiveness of k-Means Cluster based Fast k-Nearest Neighbor classification applied on Agriculture Dataset", IJCSIS, October, 2016, Vol 14, No 10.