ISSN 22773061

An Instance Selection Algorithm Based On Reverse k Nearest Neighbor

Y.Ramadevi, Y.Jagruthi, A.Sangeeta
Department of Computer Science
yrdcse.cbit@gmail.com

ABSTRACT

Classification is one of the most important data mining techniques. It belongs to supervised learning. The objective of classification is to assign class label to unlabelled data. As data is growing rapidly, handling it has become a major concern. So preprocessing should be done before classification and hence data reduction is essential. Data reduction is to extract a subset of features from a set of features of a data set. Data reduction helps in decreasing the storage requirement and increases the efficiency of classification. A way to measure data reduction is reduction rate. The main thing here is choosing representative samples to the final data set. There are many instance selection algorithms which are based on nearest neighbor decision rule (NN). These algorithms select samples on incremental strategy or decremental strategy. Both the incremental algorithms and decremental algorithms take much processing time as they iteratively scan the dataset. There is another instance selection algorithm, reverse nearest neighbor reduction (RNNR) based on the concept of reverse nearest neighbor (RNN). RNNR does not iteratively scan the data set. In this paper, we extend the RNN to RkNN and we use the concept of RNNR to RkNN. RkNN finds the query objects that has the query point as their k nearest-neighbors. Our approach utilizes the advantage of RNN and proposes to use the concept of RkNN. We have taken the dataset of theatres, hospitals and restaurants and extracted the sample set. Classification has been done on the resultant sample data set. We observe two parameters here they are classification accuracy and reduction rate.

Keywords: Reverse Nearest Neighbor, Reverse k Nearest Neighbor, reduction rate, classification.

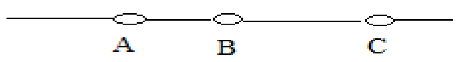
Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 10, No.7
editor@cirworld.com
www.cirworld.com, member.cirworld.com

1. INTRODUCTION

Classification is one of the most important data mining techniques which belong to supervised learning. The task of classification can be divided into two parts, training and testing. Training data is nothing but the set of records with attributes and among those attribute one is class label. A model is to be extracted from the class label attributes. Unseen data should be assigned class label. Testing is to find how accurately an unlabeled record is assigned a class label. The main aim of classification is to assign classes of unlabelled data. Classification supports in decision making, disease and text categorization etc.

With the abundant growth of data the storage capacity and classification efficiency is becoming very difficult. There may be chance of noise in the data. Therefore data reduction must be done. Data reduction is to extract subset from a data set, so the amount of data will get reduced and also noises get removed. Instance selection algorithms are used for data reduction. Instance selection algorithms choose representative samples into the subset. Representative samples are nothing but the samples which represents the entire training dataset. If noises can be removed, the accuracy of using the subset for classification is possible to be higher than using the whole dataset and also the storage efficiency also improves. Most of the instance selection algorithms use the concept of NN. But NN needs much computation time while testing because it computes all the distances between training data to assign class to unlabelled instance.

Reverse nearest neighbor reduction (RNNR) algorithm chooses the representative samples more effectively than other instance selection algorithms. It does not iteratively scan the dataset while choosing samples. It uses the concept of reverse nearest neighbor (RNN). Unlike NN, RNN set contains more than one instance. The concept of RNN is explained as: B is the nearest neighbor for both A and C, then A and C will be in the RNN set of A. In the fig.1 B is the NN of A, A is the NN of B and B is the NN of C, therefore RNN set of A contains B, RNN set of B contains A and C, but there is no instance in the RNN set of C because C is not NN to any of the two instances.



Instances	NN	RNN
A	B	B
B	A	A,C
C	B	-

Fig.1. An example of RNN

In this paper we utilize the property of RNNR instance selection algorithm for RkNN. RkNN query finds the objects that have the query point in their k nearest neighbors. RNNR applies RkNN to each class and extracts sample set from which classification will be done. In our experiments RNNR achieves high accuracy and lower reduction rate than the comparators.

2. PROBLEM DEFINITION

Classification of an instance by using the entire training data set is difficult and also may not be accurate. There may be chances of noises present in the data and also storage becomes burden. So the training data should be reduced to avoid noises and to increase the storage efficiency. For the data reduction, an instance selection algorithm called reverse nearest neighbor reduction (RNNR). RNNR utilizes the property of RNN to choose representative samples to the sample set. RNNR applies RNN to each class and selects samples which can represent other instances in the same class. While choosing representative samples noises will get eliminated.

In these experiments RNNR achieves comparable accuracy and lower reduction rate than comparators. In other words we can use smaller subset of the training data so as to obtain good classification results and lower reduction rate.

3. RELATED WORK

Data reduction methods are divided into two types, Instance Based Learning (IBL) algorithms and Clustering Based Learning (CBL) algorithms.IBL algorithms choose representative samples to sample set from the training data set. Most of the instance selection algorithms are based on the concept of nearest neighbor. Few instance selection algorithms selects samples based on two methods, incremental and decremental. Incremental algorithms select some instances as samples and iteratively add instances with different class label with their nearest neighbor to the sample set. Condensed NN (CNN), Modified CNN (MCNN), Fast NN Condensation (FCNN) etc are the examples of incremental algorithms. Decremental algorithms remove instances from the training data set with different class label with the majority of k nearest neighbor. The instances which are removed are considered as noises. That is they do not represent any instance. Edited NN (ENN), Decremental Reduction Optimization Procedure 3 (DROP3), Iterative Case Filtering (ICF) etc are the examples of decremental algorithms.

2.1 Reverse Nearest Neighbor

Reverse nearest neighbor (RNN) queries are strongly related to nearest neighbor (NN) queries. As there queries are based on the distance metric. The main difference between RNN and NN query is the target to be recorded. RNN query finds the data points that have the query point as their NN. NN and RNN are asymmetric. That is, for example a query point p has a NN q, it does not imply that p's RNN is the data point q.

For example if there are twenty five houses and if we want to open a new restaurant near those houses, before opening the new restaurant we need to consider many things whether there is another restaurant near those houses or is it really use full etc. By using the concept of RNN we can find whether the new restaurant is nearest neighbor to most of the houses. If the new restaurant is nearest neighbor to most of the houses then the RNN set of the new restaurant contains those houses for which the new restaurant is the nearest neighbor. If the RNN set of new restaurant contains many of the houses then the new restaurant can be opened at that particular location.

2.2 Reverse k Nearest Neighbor

Reverse k Nearest Neighbor (RkNN) query finds the objects that take the query object as one of their k nearest neighbors. With the wide development of location sensing devices and location based services are getting popular. Location related queries play an important role in location based services. RkNN query is one of the location related query that finds the objects whose k nearest neighbors includes the query point. For example, a taxi can issue a RkNN query to find the passengers for which the taxi is one of his/her k nearest taxis. RkNN query helps in finding the influence of query point in the data set. In location dataset, the distance between two locations p and q gives the p's influence on q and vice-versa. The shorter the distance is the higher is the influence, and an object's RkNN are the objects that are highly influenced by the query point.

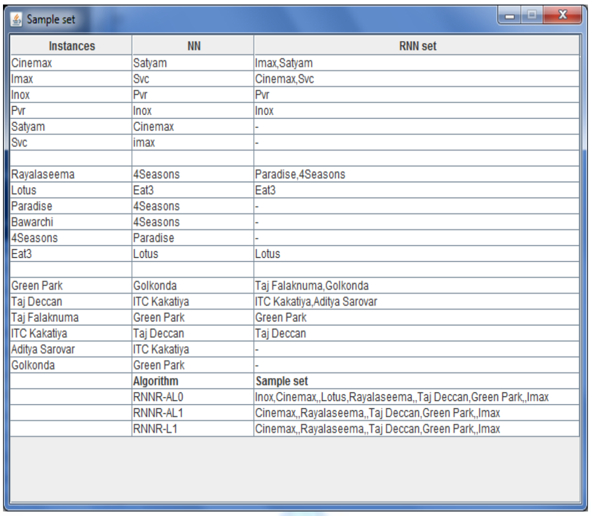
Here k is the user defined value in location based data set if we give the number of instances as six then each dataset with six instances will be displayed. If we want to extract sample set for only few instances by choosing the k value we can select the number of instances for which the sample set is to be calculated. With this the time gets saved because we no need to calculate RNN for every instance in the training dataset.

4. RESULTS

The concept of instance selection algorithm based on RkNN is implemented by taking three different datasets they are: restaurants, hotels and theatres. Final sample set is calculated from the training dataset and classification is done. Different results are shown below; the proposed paper is implemented in Java technology. The results show that the classification accuracy and the reduction rate have been achieved.

Table.1. Training dataset

Restaurants	Hotels	Theatres
Bawarchi	greenpark	Pvr
Chutneys	Taj Deccan	Imax
Eat3	Taj Falaknuma	Inox
Siaa	ITC Kakatiya	Satyam
4 seasons	Aditya Sarovar	Svc
Paradise	Athithi Inn	Big Cinemas
Lotus	golkonda	Cinemax
Rayalaseema	Delight Inn	Galaxy Theatre



Instances	NN	RNN set
Cinemax	Satyam	Imax,Satyam
Imax	Svc	Cinemax,Svc
Inox	Pvr	Pvr
Pvr	Inox	Inox
Satyam	Cinemax	-
Svc	imax	-
Rayalaseema	4Seasons	Paradise,4Seasons
Lotus	Eat3	Eat3
Paradise	4Seasons	-
Bawarchi	4Seasons	-
4Seasons	Paradise	-
Eat3	Lotus	Lotus
Green Park	Golkonda	Taj Falaknuma,Golkonda
Taj Deccan	ITC Kakatiya	ITC Kakatiya,Aditya Sarovar
Taj Falaknuma	Green Park	Green Park
ITC Kakatiya	Taj Deccan	Taj Deccan
Aditya Sarovar	ITC Kakatiya	-
Golkonda	Green Park	-
	Algorithm	**Sample set**
	RNNR-AL0	Inox,Cinemax,,Lotus,Rayalaseema,,Taj Deccan,Green Park,,Imax
	RNNR-AL1	Cinemax,,Rayalaseema,,Taj Deccan,Green Park,,Imax
	RNNR-L1	Cinemax,,Rayalaseema,,Taj Deccan,Green Park,,Imax

Fig. 2 Sample set extraction

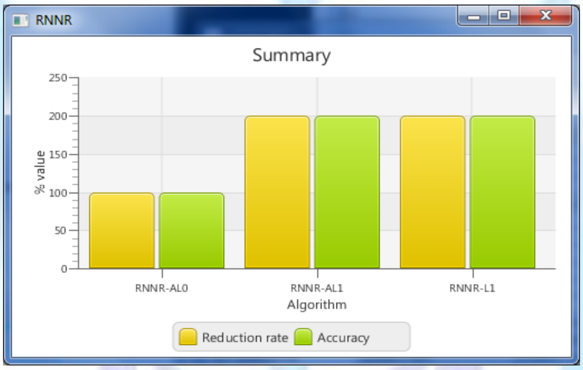


Fig.3.Comparative graph showing classification accuracy and reduction rate.

5. CONCLUSION

In this paper, a new instance selection algorithm by using the concept of Reverse k Nearest Neighbor (RkNN) has been implemented. Here k is the user defined and k value defines the number of instances in each dataset. By observing the results this concept achieves high classification accuracy and high reduction rate.

REFERENCES

[1] Bi-Ru Dai and Shu-Ming Hsu, “An Instance Selection Algorithm Based On Reverse Nearest Neighbor” PAKDD 2011, Part I, LNAI 6634, pp. 1–12, 2011.@ Springer-Verlag Berlin Heidelberg 2011.

[2] Cover, T., Hart, P.: Nearest Neighbor Pattern Classification. IEEE Trans.Information Theory 13, 21–27 (1967).

[3] Delany, S.J.: The Good, the Bad and the Incorrectly Classified: Profiling Cases for Case-Base Editing. In: McGinty, L., Wilson, D.C. (eds.) ICCBR 2009. LNCS, vol. 5650, pp. 135–149. Springer, Heidelberg (2009).

[4] Devi, F.S., Murty, M.N.: An Incremental Prototype Set Building Technique.Pattern Recognition 35(2), 505–513 (2002).

[5] Elke Achtert, Christian B¨ohm, Peer Kr¨oger, Peter Kunath, Alexey Pryakhin, Matthias Renz ” Efficient Reverse k-Nearest Neighbor Search in Arbitrary Metric Spaces” SIGMOD 2006 June 27-29, 2006, Chicago, Illinois, USA.

[6] Fabrizio Angiulli “Fast Nearest Neighbor Condensation for Large Data Sets Classification” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 11, NOVEMBER 2007.

[7] Hart, P.: The Condensed Nearest Neighbor Rule. IEEE Trans. Information Theory 14, 515–516 (1968).

[8] Tobias Emrich, Hans-Peter Kriegel, Peer Kroger, Matthias Renz, Naixin Xu, Andreas Züfle“Reverse k-Nearest Neighbor Monitoring on Mobile Objects”ACM GIS ’10, November 2–5, 2010, San Jose, California, USA.2010 ACM 978-1-4503-0428-3/10/11.

[9] Wei Wu Fei Yang, Chee-Yong Chan Kian- Lee Tan “FINCH: Evaluating Reverse k-Nearest-Neighbor Queries on Location Data” PVLDB '08, August 23-28, 2008, Auckland, New Zealand 2008 VLDB Endowment, ACM 978-1-60558-305-1/08/08.