Vol.3, No.1 ISSN:2349-0012

INNOVATIVE RESEARCH ORGANISATION

International Journal of Advance Research in Education, Technology & Management

(Scholarly Peer Review Publishing System)

A Survey of Uncertain Data Mining Techniques

Rashmi Agrawal
Research Scholar
Manav Rachna International University

Babu Ram
Manav Rachna International University

ABSTRACT

The presence of uncertainty in data affects the results of data mining techniques. The attributes which have a higher level of uncertainty needs to be treated differently as compared to the attributes having lower level of uncertainty. Different algorithms exist in literature for users to choose a suitable one as per their need. This research paper deals with the fundamentals of various existing data mining techniques for uncertain data. The data mining functionalities include classification, clustering and association rule mining.

Keywords: Uncertain data, classification, clustering, association rule, data mining

1 INTRODUCTION

Data Mining refers to extracting information or knowledge from vast amount of data. There exist various data mining analytical techniques like classification, association rule mining, clustering. Various efficient algorithms have been designed by the researchers to give efficient results using these data mining techniques to handle certain data having a fixed point definite value but Data uncertainty is common in real-world applications. Various factors like physical data generation and collection process, unreliable data transmission, transmission bandwidth, measurement errors, decision errors contribute for the uncertainty in data. The uncertainty in data may be for both numerical data and categorical data.

Basically we can divide the type of uncertainty into two categories namely existential uncertainty and value uncertainty. The first one refers to the existence of a tuple and has been handled well in the form of handling missing values by various researchers. In the second category the data item is modeled in a closed region with multiple possible values. According to Cheng and Prabhakar [1], the notion of probabilistic answers to queries over uncertain data has not been so far well studied. From the uncertain data providing meaningful answers is a tedious exercise. However, we can observe that in many applications, the values of objects cannot change drastically in a short period of time and the rate of change or degree of varies of an object may be constrained. For example, the temperature recorded in a day may not change more than 5 degree in 2 hours. By using such kind of information we can record a range of value as maximum and minimum. Cheng [1] has provided a broad classification of probabilistic queries over uncertain data.

The aim of this paper is to discuss various data mining techniques for handling uncertain data.

2 MEASURES OF UNCERTAINTY

It has been shown by the researchers that there exists various types of uncertainty in the data and different models can be used to represent that uncertainty. In this section we present a brief of these uncertain measures.

2.1 Shannon's Measure of Uncertainty

The famous uncertainty model of Shannon is known as Shannon's entropy. Assume there are N data points and any Kth data vector is {Ckj}, j=1…N with probability of {p(Ckj)}. Then Shannon's entropy HS , of the probability vector, { p(Ckj) }, is:

$$H_S(P_k) = H_S^{P_k} = -\sum_j p_{kj} \log_2 p_{kj}$$

Where Pkj= p(Ckj) = Probability that Ckj will occur and Pk=p(Ckj) is the probability of Kth data point.

INNOVATIVE RESEARCH ORGANISATION

International Journal of Advance Research in Education, Technology & Management

(Scholarly Peer Review Publishing System)

2.2 Measures of Probabilistic Uncertainty

Measurement uncertainty is associated with its standard deviation, and can be represented by positive square root of the variance. This is directly related to mean width of the distribution i.e. the probability density functions. Any useful Probability density function can be applied eg. Gaussian function.

2.3 Measures of Fuzzy Uncertainty

A proper measure for fuzzy uncertainty satisfies the five properties Sharpness, Maximality, Resolution, Symmetry and Valuation. Shannon also represented a fuzzy version of functional entropy.

3 DATA MINING TECHNIQUES

According to Han and Kamber [2], data mining functionalities include classification, clustering, frequent pattern mining and association rule analysis. Classification technique is used to classify the data into the specified classes. It is a supervised learning technique that induces a classification model from a database. Clustering technique analyzes data objects without consulting a known class model. Clustering is an unsupervised learning technique. Association rule mining helps in finding the interesting patterns and correlations.

3.1 Classification

Classification is one of the important techniques in data mining which is used to classify the data into the specified classes. In this technique, a training set is used where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model, used to classify new objects.

A decision tree is a simple recursive structure for expressing a sequential classification process in which a case described by a set of attributes is assigned to one of a disjoint set of classes. Various efficient algorithms have been developed to construct a decision tree in a reasonable amount of time. Generally, these algorithms (ID3, C4.5, CART, SPRINT) follow the greedy approach by taking the local optimum decision. Usually information gain, entropy, gini index and other similar measures are used to take the decision of selecting the attribute in split point. Assume there are two classes, *P* and *N* and let the set S of examples contain *p* elements of class *P* and *n* elements of class *N* then the information gain can be calculated as

The classifier selects the attribute with the highest information gain and a decision tree is build. These traditional algorithms works well on certain data but cannot be applied on uncertain data.

Biao Qin *et al.* [3] presented a decision tree based classification and prediction system for uncertain data. This tool used measures for constructing, pruning and optimizing decision tree. The measures were computed considering uncertain data probability distribution functions. Based on the measures, the optimal splitting attributes and splitting values were identified and used in the decision tree. The author have shown that DTU was capable of processing various types of uncertainties and has satisfactory classification performance even when data was highly uncertain.

Tsang, Kao [4] gives classification results as a distribution for each test tuple, giving a distribution telling how likely it belongs to each class. In their model, a dataset consists of *training d tuples*, {t1, t2, . . . , td}, and k numerical (real-valued) feature attributes, A1, . . .Ak. Each tuple ti is associated with a feature vector Vi = (fi,1, fi,2, . . . , fi,k) and a class label ci ∈ C. Here, each fi,j is a pdf with domain [ai,j , bi,j] modelling the uncertain value of attribute Aj in tuple ti. The classification problem is to construct a model M that maps each feature vector (fx,1, . . . , fx,k) to a probability distribution Px on C.

K Nearest Neighbour classifier uses a simple classification technique. It identifies a sample of the same class that lies in its closest K instance. This distance can be measured by various distance functions like Euclidean and Manhattan. Of course, these distance functions are used only for continuous variables.

INNOVATIVE RESEARCH ORGANISATION

International Journal of Advance Research in Education, Technology & Management

(Scholarly Peer Review Publishing System)

For the categorical variable hamming distance is used to measure the distance between objects. To compute the distance between two objects the following Euclidean distance may be calculated

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

Richard J. Samworth [5] has derived an asymptotic expansion for the excess risk of a weighted nearest neighbor classifier. He has shown that the most obvious defect with the nearest neighbour classifier is that it places an equal weight on the class labels of each of the k nearest neighbour to the point x, being classified. He suggested improvements in the misclassification rate by putting decreasing weights on the class labels of the successively more distant neighbors.

Elena Marchiori [6] investigated a decomposition of RELIEF (an existing feature weighting algorithm) into class dependent feature weight vectors, where each vector describes the relevance of features conditioned to one class. The results of experiments indicated the usefulness of this decomposition for unraveling relevant features for a single class and for providing information about the difficulty of the considered learning task.

Angiulli and Fassetti [7] introduced an uncertain nearest neighbor rule generalizing the certain nearest neighbor rule to uncertain scenario. They assumed that an uncertain object is an object whose actual value is modeled by a multivariate probability density function. The Uncertain Nearest Neighbour rule (UNN) relies on the concept of nearest neighbour class rather than nearest neighbour object.

3.2 Clustering

A clustering technique group the objects into k clusters by minimizing the total expected distance from the objects to a representative point of their corresponding clusters. Due to the presence of uncertainty the behavior of underlying clusters is affected which can significantly affect the quality of the result.

K-Means algorithm organizes objects into k partitions, where each partition represents a cluster. If the cluster mean is Cj, the minimized sum of squared errors (SSE) can be calculated by using the formula-

$$\sum_{j=1}^{K} \sum_{x_i \in C_j} \left\| \mathbf{c}_j - \mathbf{x}_i \right\|^2 ,$$

Where $\| \cdot \|$ is a distance metric between a data point $\mathbf{x}i$ and a cluster mean $\mathbf{c}j$.

To consider the uncertainty in the clustering, Chau, M., Cheng [8] proposed a UK means clustering algorithm where data object xi is specified by an uncertainty region with an uncertainty pdf f(xi). Given a set of clusters, Cj's the expected SSE can be calculated as follows:

$$= \sum_{j=1}^{k} \sum_{i \in C_j} \int \left\| \mathbf{c}_j - \mathbf{x}_i \right\|^2 f(\mathbf{x}_i) d\mathbf{x}_i$$

UK-means compute the *expected* distance and cluster centroids based on the data uncertainty model.

A popular method for clustering of deterministic data is Mixture model clustering Dempster, Laird [9], which models the clusters based on number of probabilistic parameters. A popular algorithm for this is EM algorithm. Hamdan, and Govaert [10] Generalized this algorithm for uncertain data. In the modified algorithm each data value may be drawn from an interval. The difference between the uncertain and basic version of the algorithm is that each instantiation is treated as an uncertain value of the record rather than a fixed value. To evaluate the expressions in the E-step and M-step the EM algorithm is also changed. The researcher pointed out that the approach can be used fairly easily for any uncertain distribution represented in the form of a probability histogram of values.

Another category in clustering is density based method in which clusters are regarded as regions in data space in which the objects are dense. The key point in density based clustering is that Density based clustering sees the cluster as the high density area which is further divided by low density area. The En-DBSCAN algorithm combines the characters of uncertain data and improved the DBSCAN algorithm

INNOVATIVE RESEARCH ORGANISATION

International Journal of Advance Research in Education, Technology & Management

(Scholarly Peer Review Publishing System)

in two areas. It uses the probability radius concept which can change the length of center point's cluster radius according to the probability. Further this method uses entropy to measure the uncertainty of the object.

3.3 Association Rule Mining

Association rule mining is one of the best researched techniques of data mining. The goal of association rule mining is to find interesting correlations, frequent patterns or casual structures in the set of items in the database. Association rules are used in various areas such as market and risk management, risk control and telecommunications etc. . . .

An association rule [11] is an implication in the form of X→ Y , where X, Y $\subset$ I are sets of items called item sets, and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent. Two important measures for association rule are support and confidence. Support(s) of an association rule is defined as the percentage/fraction of records that contain X$\cup$Y to the total number of records in the database.

$$Support(XY) = \frac{Support\ count\ of\ XY}{Total\ number\ of\ transaction\ in\ D}$$

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X $\cup$Y to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule X→Y can be generated.

$$Confidence(X|Y) = \frac{Support(XY)}{Support(X)}$$

Agarwal et al in 1993 proposed the first algorithm RIS for association rule mining. In this algorithm only one item consequent association rules are generated. Apriori algorithm is considered as a great algorithm in the association rule mining. Apriori algorithm was first proposed by Agrawal and Srikant in 1994. Apriori algorithm also have the drawback of scanning the whole data bases many times. Based on Apriori algorithm, many new algorithms were designed with improvements. Although these algorithms works fairly well with certain data but does not give satisfactory results with uncertain data.

Chui, Kao, & Hung [14] have adapted the basic Apriori algorithm for uncertain data and named it as U – Apriori. They consider transactions whose items are associated with existential probabilities and give a formal definition of frequent patterns under such an uncertain data model. To improve mining efficiency a data trimming framework was proposed to reduce the database by removing items with low probability. This mining process was very low and not efficient in space.

Chui, C.-K., & Kao [15] developed a technique called decremental pruning to improve the performance of U-Apriori. In this technique the database is scanned and counter is updated by subtracting the corresponding "over-estimate" for each item in the pattern. It improves the efficiency of U-Apriori in terms of run time and memory requirements.

An efficient tree-based algorithm for mining uncertain data was proposed by Leung, C. K.-S., Mateo [16]. They developed UF tree which was a variant of FP tree. When compared with U-Apriori, the proposed UF-growth algorithm was found to be superior in performance.

4 CONCLUSION

In terms of performance the algorithms developed so far for precise data in different data mining techniques like classification, clustering and association rule mining, we get satisfactory results but the uncertain data provides a completely different scenario and most algorithms give different results when applied on this data. In this paper we have studied few techniques of data mining algorithms on uncertain data. Uncertain data mining is an area of interest for researchers and much work is required to handle the uncertain data in a better way. We further plan to go into the detail of classification technique for uncertain data.

INNOVATIVE RESEARCH ORGANISATION

International Journal of Advance Research in Education, Technology & Management

(Scholarly Peer Review Publishing System)

5 REFERENCES

[1] Reynold Cheng Dmitri V. Kalashnikov Sunil Prabhakar 2003, Evaluating Probabilistic Queries over Imprecise Data, Proceedings of the 2003 ACM SIGMOD international conference on Management of data, ACM New York, PP 551-562,

[2] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, London: Academic Press, 5, 2001.

[3] Biao Qin,Yuni Xia,Rakesh Sathyesh,Jiaqi Ge,Sunil Probhakar, 2011, Classify Uncertain Data with Decision Tree, Database Systems for Advanced Applications Lecture Notes in Computer Science Volume 6588, pp 454-457,.

[4] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, Sau Dan Lee 2011 Decision Trees for Uncertain Data, IEEE transaction knowledge and data engineering, vol. 23 no. 1, pp. 64-78, .

[5] Richard J. Samworth 2012 Optimal Weighted Nearest Neighbour Classifiers, The Annals of Statistics, Vol. 40, No. 5, 2733-2763.

[6] Elena Marchiori 2013 Class Dependent Feature Weighting and K-Nearest Neighbor Classification, Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science Vol. 7986, pp 69-78.

[7] Fabrizio Angiulli, Fabio Fassetti 2013 Nearest Neighbor-Based Classification of Uncertain Data, Journal ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 7,No.1.

[8] Chau, M., Cheng, R., and Kao, B., 2005 Uncertain Data Mining: A New Research Direction, in Proceedings of the Workshop on the Sciences of the Artificial, Hualien, Taiwan, December 7-8.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin.1977 Maximum likelihood for incomplete data via the EM algorithm., Journal of the Royal Statistical Society, Series B, 39: pp. 1–38.

[10] H. Hamdan, and G. Govaert.2005 Mixture model clustering of uncertain data, In Proceedings of IEEE ICFS Conference, pp. 879–884.

[11] Technical Report,2003 CAIS, Nanyang Technological University, Singapore, No. 2003116 .

[12] Agrawal, R., Imielinski, T., and Swami, A. N.(1993) Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, . Washington, D.C., 207.

[13] C. E. Shannon and W. Weaver,.1949 The Mathematical Theory of communication, Urbana, The University of Illinois Press.

[14] Chui, C.-K., Kao, B., & Hung, E. 2007 Mining frequent itemsets from uncertain data, Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), (pp. 47–58). Nanjing, China.

[15] Chui, C.-K., & Kao, B. 2008, A decremental approach for mining frequent itemsets from uncertain data, Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), (pp. 64–75). Osaka, Japan.

[16] Leung, C. K.-S., Mateo, M. A., & Brajczuk, D. A.2008 A tree-based approach for frequent pattern mining from uncertain data, Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), (pp. 653–661). Osaka, Japan.