Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 3, March-2017

Issues in Data mining: An Ample Review

Abhishek A. Gulhane[1], Ashwini A. Gulhane[2], Smeet D. Thakur[3], Rupesh M. Hushangabade[4], Nilesh S. Wadhe[5]

[1]*Information Technology Department, PRMIT & R, Badnera*
[2]*Electronics & Telecommunication Department.*
[3] *Information Technology Department, PRMIT & R, Badnera*
[4 5] *Information Technology Department, PRMIT & R, Badnera*

Abstract —Data mining has attained stunning success in almost every province such as wireless sensor network, social network, health care etc with expansion of its various algorithms. Every data mining algorithm has its natural boundaries. The application domain and the actual data, both together, heavily sway the particular choice as well as recital of any data mining, machine learning or statistical algorithm. The role that this review makes is that it detailed a number of data mining issues along with the metrics to compute the data quality and algorithm performance under a single hood. This paper has explained most critical issues in data mining, i.e., Missing Value citation, Phase collection, Outlier revealing, Cluster examination of high dimensional data, Extreme classes in classification, Privacy of data, mining from complex/distributed data. It not only presents these issues but also discusses their existing solutions. Survey also throws light on the boundaries and research breach for forthcoming researchers. This ample understanding of the issues and metrics can be a luxury for beginners in the research of data mining. Survey shows that the most frequently used algorithm performance measures are accuracy and time complexity.

Keywords- Disparity classification; clustering; privacy data mining; Missing value citation (MVC); Data mining.

I. INTRODUCTION

With the encroachment of knowledge in terms of software, hardware, and automation of business enterprises, a great deal of data has been engendered and gathered in data marts and data warehouses. As an outcome, conventional and informal blend of numerical techniques and data management tools are no longer adequate for study of this gigantic set of data [1]. Hence, an bright data study technique such as Knowledge discovery has been discovered so that knowledge could be extort from a variety of databases. Data mining is one of the steps of knowledge discovery. It is used to pull out the desired information from given data. It could be applied to any field such as healthcare, bank, science and others so that we are able to envisage result and expose patterns in data. There are several troubles in the data mining. To handle these problems, researchers have projected several methods and evaluated their results using different metrics. In this work, we first have studied various concerns in the field of data mining research. Second, we classified performance metrics used by researchers to confirm their approaches to engage in different data mining issues. Third, we have abridged recital/eminence actions showing the referred papers and analyzed the most popularly used measures. Rest of the paper is divided into four sections. Section II promulgates issues with ongoing research along with their research breach. Section III categorizes the performance measures according to data mining issues. These recital/eminence actions are used to validate the work while dealing with a variety of issues. Section IV concludes the paper.

II. ISSUES IN DATA MINING

Data mining issues with their enduring research are conversed in following segment.

A. Missing Value Citation

In view of the fact that many real world data sets contain missing values (MVs), acquiring superiority mining results from such inadequate data may be very rigid. Typically, the management of MVs in data mining can be carried out in three distinct ways [2]. Easiest way is to get rid of the occurrences that are having MVs in their feature. Hence, eradicating columns where the number of missing values is high is incorporated in this type too. An additional way estimates the constraint such as variation and covariance based on complete data using maximum probability procedures [3], and then uses these constraints for citation. Finally, the citation of missing values consists of schedules that are used to fill in the missing values with estimated ones. When the attributes of data set are dependent on each other, MVs can be calculated by influential the associations among element. Plentiful loom have been studied in MV prose that portray that how to pact with missing values [3, 4].Broadly they can be classified into different groups such as list wise deletion, mean/mode substitution, maximum probability methods, multiple citations and machine learning based citations [5]. Recently [4] has wished-for an approach for mixed data set consisting of distinct as well as incessant attributes. Moreover, a novel clustering based multiple citations [6] via Grey relational analysis [5] as well as Auto adaptive imputation [49] has also been projected.

Research Breach: Different loom for missing value citation discussed above are employed to prepare data in order to avoid the negative impact of incomplete data on mining tasks. But most of these approaches have certain boundaries. Firstly, most of them are based on hypothesis that the various attributes of data records are sovereign of each other. However, real world data may have inherent link across multiple attributes. This may direct to incorrect filling of

All Rights Reserved, @IJAREST-2017

missing values. Secondly, most missing value citation methods are applicable on numeric data only. But the real world data may be numeric, unconditional or mixed type. In such cases, most methods fail. However, an important issue like noise has been uncared for in current prose that may be present in the data set in addition to missing values. Hence impact of noise in implicating missing values should be investigated in more MV approaches. Further, MV approaches should also consider context information of nearest neighbors from which the missing value is assigned in most of approaches.

B. Phase collection

Dataset used for mining of data may consist of numerous scopes but all of them may not be appropriate. Hence, aspect collection removes distinct, abandoned, or noisy data and reduces the number of columns that results into better feat of classification, alliance and association algorithms. Feature selection process consists of of four steps that is feature rift collection, subset assessment, stopping criteria and validation [7]. The subset selection is influenced by the search space such as random, complete or sequential. Features can be chosen either in forward or backward manner from the given data set. The most relevant feature subset is evaluated using some assessment parameters such as deviation, probability of error, doubt, reliability, trust, and interclass distance. Feature selection approaches are largely classify into filter [8], wrapper [9] .and hybrid [10]. Feature selection algorithms have also used soft computing techniques such as genetic algorithm [11] and neural networks. Recently, worldwide and local structure protection framework for based feature selection algorithm [50] is proposed that gives the importance to local geometrical structure of the data.

Research breach: More research is required when it becomes very costly to trace bulky data sets in several periods. Offered algorithms find the difficulty to search the occurrence at random in case of continuous data. Recently, mixture of wrapper and filter approaches are considered to handle data set with high dimensionality. These algorithms results into best possible performance with a particular mining algorithm with similar time convolution of filter algorithms. Hence, more skillful (i.e. computational power, statistical accuracy and algorithmic stability) search approaches and evaluation criteria are required to select features from a given data sets that consist of millions of dimensions. Further, there is a need to exploit label correlations for effective feature reduction in high dimensional multi-label learning. Moreover, feature selection in presence of noisy data is still an open issue.

C. Outlier revealing

Data sets used for data examination may consist of data occurrence that does not obey the rules to the general properties of the other data occurrence present in the data sets. These objects are recognized as outliers or anomalies. The study of outlier data is known as outlier mining. Traditional methods to detect anomalies have been grouped as type-1(unsupervised), type-2 (supervised), type-3 (semi supervised) in the outlier literature. Most widespread methodologies for anomaly detection are statistical models, neural network forms, machine learning models like SVM, and hybrid forms. Statistical models are suitable for the data sets in which ordinal data can be converted to numerical values so that data set can be processed statistically. Otherwise, complex transformation results into increased processing time .Therefore, approaches based on neural network were proposed .But these approaches were subjected to the problem of "Curse of Dimensionality". Moreover, both the approaches need the data that is either cardinal or at the least ordinal so that vector distances could be calculated. If categorical data is not ordered, these approaches cannot be applied. Machine learning approach like decision tree C4.5 is able to detect anomalies in categorical data. Therefore, it identifies errors and undesired entries in databases. Fusion model such as MLP with a Parzen window novelty recognizer [12] that uses MLP with Gaussian kernel is a grouping of two methodologies. Recently subspace outlier detection algorithms [13] like ENCLUS,PODM and HiCS [14] have been argue in the literature that are not curbed to one subspace .These methods detect subspaces with low entropy and high interest. [13] also proposes a new measure Commutative Mutual Information (CMI) that groups features into different subspaces according to strong mutual dependency.

Research breach: Subspace based outlier detection techniques such as ENCLUS, PODM discretize continuous features to calculate entropy measures. It results into loss of knowledge. Although HiCS does not convert the continuous features; it loses useful subspaces due to its random nature. Moreover, mentioned algorithms are not suitable for high dimensions. It necessitates more algorithms to deal with high dimensions. Further outlier detection techniques need to be investigated for spatial, temporal data and video data.

D. Cluster examination of high dimensional data

Basic clustering methods [15] are hierarchical, partitioning based, grid-based, density based, and model based approaches. These approaches are however not optimal for computational biology and clinical data which are high dimensional. As dimensions of the data grow, for any point the difference between the distance to its nearest point and that to the farthest point turn out to be insignificant. This phenomenon may leave the results of clustering algorithm sensitive to any small perturbation to the data due to noise and make the exercise of clustering useless [16]. Although feature selection/dimension reduction algorithms reduce high dimensional data by choosing some attributes, it cannot be applied to clustering of high dimensional data. In clustering, many clusters may be present in different subspaces of smaller dimensions. But the sets of dimensions may be overlapping or non overlapping [7]. Common Approaches to deal with this problem are Subspace clustering. Several algorithms [17] such as CLIQUE, ENCLUS, MAFIA, and SUBCLU have been propounded to get the subspace clusters. Recently density conscious subspace clustering has been proposed by [17] that group the instances according to the relative region densities of the subspaces instead of using density threshold. Existing clustering algorithms are not appropriate for the situations when the relationship between data points changes

All Rights Reserved, @IJAREST-2017

with the time. It is unreasonable to recluster such large scale data sets. To handle this type of situation, [18] proposed evolutionary clustering with low rank kernel factorization that partitions data at every time step. [19] also discusses Fuzzy approach for Multi-type Relational clustering that clusters four type of objects. Recently, actionable subspace clustering [48] is proposed that incorporates the domain knowledge and suggests the actions to be taken.

Research Gaps: Mining the data from different types of objects is essential especially when data are related to each other A little work like [19] has developed an approach that does fuzzy clustering of relational data that involve several object types**.** Hence, more approaches need to be investigated. Although several algorithms have been developed in the literature of subspace clustering, these are based on the assumption on perfect data as well as correct knowledge of number of subspaces. However, few algorithms like generalized principal component analysis (GPCA) [20] has considered the unknown number of linear subspaces. But it has not considered the noisy data. Hence, correctness of existing algorithms in presence of noise needs to be investigated. Moreover, getting clusters from multiple non linear subspaces in presence of noise still needs to be addressed by researchers. Further, more actionable subspace clustering algorithm needs to be investigated for different kind of applications.

E. Extreme classes in classification

A class-imbalanced classifier is a rule to forecast the class members of new samples from an available dataset where the class sizes disagree substantially. When the class sizes are very distinct, most benchmark classification algorithms may favor the bigger (majority) class producing in poor accuracy in the minority class proposition. Numerous approaches have been suggested to the problem of imbalanced class in two groups namely data as well as algorithm In data group, several sampling methods like query based learning sampling [21] as well as clustering based pre-processing methods such as SMOTE [22] are developed. [23] has described two methods to deal with class imbalance problem. In the first method, cluster based over-sampling is done or, majority class is clustered first. In second method, clustering each class identifies subclusters. Later on, re-sampling of each sub-cluster is done to increase class-size. This removes the class imbalance. The treatment of the imbalanced problem at the data level by means of preprocessing techniques such as SMOTE [22] has proven to be very useful, and it has the advantage that there is no need to make any changes to the classification algorithms. [22] has insisted on class imbalance ratio(IR) for better classification accuracy . In the algorithmic group, approaches include the costsensitive adaptations to the different classes of the problem, or accept variations of the likelihood estimation in decision trees [24], and decision threshold adaptation. [25, 26] have modified the SVM algorithm that re-arranges the limits for the cases where it is imbalanced. Cost sensitive learning sampling approaches [27] combines the approaches of data as well as algorithm groups.

Research Gaps: Approaches under data level have taken imbalanced ratio (IR) for better classification accuracy but there are still other measures such as class overlapping and dataset shift problems that may be present in the data .These may lead to inaccurate classification results. Hence, in future algorithms should be designed to address these problems. Although little work like fuzzy support vector machine for

class imbalance problem has considered noise .Still, there is a need to investigate the methods for class imbalanced data set for large scale classification problems for noisy data. Few work such as dynamic sampling for MLP [46], AdaBoost.NC [47] has been done for multiclass problems, hence more efficient work need to be done. Further, imbalanced data classification for multiple classes needs to be explored.

F. Privacy of data

Although data mining is possibly useful, numerous data holders are reluctant to supply their data for data mining for the fear of violating one-by-one privacy. In recent years, study has been made to double-check that the sensitive data of individuals will not be identified easily. A number of techniques such as randomization, k-anonymity [28], and ldiversity have been proposed in alignment to present privacypreserving data mining. The randomization method utilizes data distortion methods namely additive as well as multiplicative perturbations. k-anonymity [28, 45] is another approach that stops joining attacks by generalizing and/or suppressing attributes of data. This method is productive in preventing identification of a record. But, it may not always be effective in preventing inference of the perceptive values of the attributes of that record. Therefore, the method of l-diversity was proposed which not only sustains the lesser assembly dimensions of L, but also focuses on sustaining the variety of the crucial attributes. Numerous data repositories may be accessed via a public interface that permits cumulative querying. This gives a chance to a smart opponent to find perceptive details of the data using a chain of queries. This kind of inference may lead to full disclosure, in which an opponent may find the exact details of the desired features. A second idea is that of partial disclosure in which the opponent may be adept to slender down the standards to a variety, but may not be capable of finding accurate worth. The fundamental advances are query auditing and query inference control [29, 27]. Mostly data is distributed across various sites or computers. A single site may require the information from different data sets which are partitioned over these sites either horizontally or vertically. While the one-by-one sites may not desire to share their entire data sets, they may allow restricted data sharing with the use of a kind of protocols such as Secure Multiparty Computation [28](SMC) that only shares the data mining results among multiple parties owners. SMC consists of various operations like secure sum, set intersection, scalar product, and set union. Recently Lunning [30] has proposed a new method over secure sum for sharing simple integer data that is more efficient. The outcomes of a classification problem consist of sensitive data from the perspective of data holder. Thus the issue is how to make alterations in the data so that the correctness of the classification outcomes is decreased. At the same time, usefulness of the data for other associated applications should not

All Rights Reserved, @IJAREST-2017

be lost. A great deal of methods such as parsimonious downgrading [31] has been proposed in the literature to maintain the privacy in classified data. facilitate association rules competently. Such rules depict essential goal trading data about a business. This confidential information is a threat to the business. This gives rise to association rule hiding [32]. Two broad advances in the literature for hiding are Distortion & Blocking [33]. But when we apply these two methods, some inadequate rules may be vanished along with perceptive directions, and new phantom rules may be conceived as we deform or barricade the process. These issues diminish the utility of the data used for data mining, hence it is undesirable.

Research breach: Although SMC has been proposed for sharing results among multiple parties, there is need for more approaches that are more efficient. Further, large dimensionality and continuous stream of data also poses challenges in the research community. Moreover, approaches are to be investigated to cope up with privacy threatening technology such as facebook and Radio-frequency identification (RFID) where data is tremendously growing with the time as well as space.

G. Mining from complex/distributed data

The function of data mining in enterprise, research, surgery and other areas has been quickly growing This has resulted in the emergence of Complex data such as World Wide Web pages, DNA representations, sequential and high dimensional structures, etc that may involve heterogeneous data sources such as data is stored in various relational tables/ databases located at distinct locations. Hence traditional data mining methods cannot be applied because they assume that data is stored in one file. It has led to development of three approaches namely multirelational data mining, multidatabase mining and combined mining [34] .Multirelational data mining as well as multidatabase mining techniques extract the important features from various tables into single table. Then pattern is found by combining features from various tables. [35] has developed CrossMine-tree and CrossMine-Rule classification techniques in multirelational databases. Combined data mining does not join the related tables, instead finds the clusters of patterns from multiple data sets.

Research Gaps: Various enterprise applications such as health care, traffic surveillance etc. involve the complex data. It consists of heterogonous features namely user's demographic factors, preferences, behavior etc data which is distributed at multiple data sources. Hence there is a need to find the informative patterns from this kind of data. Moreover, many sophisticated visualization concepts such as Parallel Coordinates, RadViz or Glyphs and others discussed in [36] have been developed, but in business enterprise hospital information systems they are still not in use.

III. ANALYSIS & RESULTS

We have examine 50 papers for our study and found that authors are taking different evaluation principle to assess their proposed method. Table [1] below shows various evaluation metrics used in papers to authenticate various approaches presented in the papers. The most frequently used performance measures are accuracy and time complexity that have been used by the researchers in the papers. It shows that researchers must mention time as well as correctness taken for their algorithms so that it could be compared with previous methods.

All Rights Reserved, @IJAREST-2017

TABLE 1. PERFORMANCE MEASURES FOUND IN PAPERS

Performance measures	*Papers*
Risk Ratio , Leverage, Potential Casual Leverage , Coverage	[38]
CMI(Cumulative mutual Information)	[13]
ROC (Receiver Operating Characteristics)	[39],[40], [10], [12]
AUC (Area under Curve)	[39], [10],[26], [12],[14], [13], [46]
Sensitivity	[25], [10]
Specificity	[39], [10],[25]
Accuracy	[15],[41], [40],[42],[18],[19],[37], [10],[23],[25],[24],[43],[33\],[5],[2],[50] ,[48]
Confusion	[15]
Coverage	[15]
Support	[38], [43], [45]
Confidence	[38], [15],[43], [45]
Kappa's Rule, Rulebase Size	[15]
Wilcoxon Signed Rank test	[3], [42]
Wilson's Noise	[3], [42], [7]
MI(Mutual Information) ratio	[3], [42], [18], [7], [19]
RMSE	[4],[41], [42], [7],[5]
Co-relation Coefficient	[4]
Time Complexity	[40],[17], [18], [19],[29],[28],[2], [15], [7],[12], [14],[13],[21],[22],[43]
Variance	[18], [37]
Recall	[17],[43], [47]
Precision	[17], [47]
Density Ratio	[17]
Space Complexity	[41], [18]
Concept Complexity	[23]
Median	[27]
Differential Entropy	[33]
NMAE(Normalized Mean Absolute Error)	[44]

IV. CONCLUSION

Data mining has been an active field of research for a long time. But before applying data mining techniques such clustering, classification and association, one has to preprocess the data to get a good quality results. Hence this paper has discussed core issues found in data mining community with their ongoing research. Our study raises the research gaps for which research community is still working or yet to be done. It also classifies performance/ quality measures required for validation of research work for specific data mining issue .This paper has also analyzed a total of 50 papers from the perspective of performance measures that have been used by the authors to verify their outcomes related to specific data mining issue. Finally, it propounds the popularly used performance measures are time complexity as well as accuracy that need to be calculated for every novel algorithm proposed by researchers.

REFERENCES

[1] M. Sushmita, P. K. Sankar, M. Pabitra , "Data Mining in Soft Computing Framework: A Survey," IEEE Transaction on Neural Networks, vol. 13, no. 1, pp, 3-14, Jan 2002.
[2] A. Farhangfar , L.A. Kurgan, and W. Pedrycz., "A novel framework for imputation of missing values in databases," IEEE Trans Syst Man Cybern Part A 37(5), pp. 692–709, 2007
[3] J. Luengo, S. García, and F. Herrera,"On the choice of the best imputation methods for missing values considering three groups of classification methods," Journal of Knowledge Information Systems, vol. 32, pp. 77–108, June 2011.
[4] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing Value Estimation for Mixed-Attribute Data Sets," IEEE Transaction on Knowledge and Data Engineering, vol. 23, no. 1, pp. 110-121, Jan 2011.
[5] J. Tian, B. Yu, D. Yu, and S. Ma, "Clustering-Based Multiple Imputation via Gray Relational Analysis for Missing Data and Its Application to Aerospace Field," The Scientific World Journal, 10 pages, 2013.
[6] T. E. Raghunathan, J. P. Reiter, and D. B. "Multiple imputation for statistical disclosure limitation," Journal of Official Statistics, vol.19, pp. 1-16, 2003.
[7] H. Liu, and L.Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, No. 4, pp. 491-502, April 2005
[8] L. Yu and H. Liu, "Feature selection for high-dimensional data: a fast correlation-based filter solution," 20th International Conferences on Machine Learning, Washington, DC, 2003.
[9] R. Kohavi and G.H. John, "Wrappers for feature subset selection," Artificial. Inteligence, vol. 97 (1–2), pp. 273–324, 1997.

All Rights Reserved, @IJAREST-2017

[10] Y.Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification", Journal of Biomedical Informatics, vol. 43, pp. 15–23, 2010
[11] C. Huang and C.Wang, "A GA-based feature selection and parameters optimization for support vector machines," Expert Systems with Applications vol. 31, pp. 231–240, 2006.
[12] V. J. Hodge & J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Review vol. 22, pp.85–126, 2004.[13] H. V. Nguyen, E. Muller, J. Vreeken, F.Keller, and K. Bohm, "
[13]CMI: An Information-Theoretic Contrast Measure for Enhancing Subspace Cluster and Outlier Detection," Proceedings of the SIAM International Conference on Data Mining (SDM), Austin, Texas, USA, 2013.
[14] F. Keller, E. Müller, and K. Böhm," HiCS: High Contrast Subspaces for Density-Based Outlier Ranking," ICDE, pp. 1037-1048, 2012.
[15] J. Han and M. Kamber, "Data Mining Concepts and Techniques," 2nd ed. Morgan Kaufmann Publishers, 2006.
[16] H. Kriegel, P. Kroger, and A. Zimek, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," ACM Transactions on Knowledge Discovery from Data, , vol. 3, No. 1, March 2009.
[17] Y.H. Chu, J.Huang, K.T. Chuang, "Density Conscious Subspace Clustering for High-Dimensional Data, IEEE Transactions on Knowledge and Data Engineering, vol. 22, No. 1, January 2010
[18] L. Wang et al., "Low-Rank Kernel Matrix Factorization for Large-Scale Evolutionary Clustering", IEEE Transactions on knowledge and Data Engineering, vol. 24, No. 6, June 2012
[19 J. Mei and L. Chen, "A Fuzzy Approach for Multitype Relational Data Clustering", IEEE Transactions on Fuzzy Systems, vol. 20, No. 2, April 2012.
[20] R. Vidal, Y. Ma, and S. Sastry, "Generalized Principal Component Analysis (GPCA)," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1–15, 2005.
[21] N. Abe, "Sampling approaches to learning from imbalanced datasets: active learning, cost sensitive learning and beyond," ICML-KDD'2003 Workshop: Learning from Imbalanced Data Sets, 2003.
[22] V. López, A. Farnandez, J.M. Torres, and F. Herrera et al, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics", Expert Systems with Applications, vol 39, pp. 6585–6608, 2012.
[23] T. Jo and N. Japkowicz, "Class Imbalances versus Small Disjuncts ", SIGKDD Explorations 6(1), June 2004.
[24] G. M. Weiss and F.Provost ," Learning when training data are costly: The effect of class distribution on tree induction," Journal of Artificial Intelligence Research, vol. 19, pp. 315–354. October 2003.
[25] G. Wu E. Y. Chang,"Class-Boundary Alignment for Imbalanced Dataset Learning," ICML-KDD'2003 Workshop: Learning from Imbalanced Data Sets, 2003.
[26] B. Raskutti and A.Kowalcyzk, "Extreme Re-balancing for SVM's: a case study," ICML-KDD'2003 Workshop: Learning from Imbalanced Data Sets, 2003.
[27] D. Dobkin., A. Jones, and R. Lipton, "Secure Databases: Protection against User Influence, "ACM Transaction on Databases Systems, vol. 4(1), 1979.
[28] A. A. Hussien1, N. Hamza2, and H. A. Hefny , "Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing , Journal of Information Security, vol. 4, pp. 101-112, 2013.
[29] I. Dinur and K. Nissim.,"Revealing Information while preserving privacy," ACM PODS Conference, 2003.
[30] L. A. Dunning, and Ray Kresman "Privacy Preserving Data Sharing With Anonymous ID Assignment," IEEE Transaction on Information Forensics and Security, vol. 8, no. 2, February 2013.
[31] L. Chang and I. Moskowitz, "Parsimonious downgrading and decision trees applied to the inference problem", New Security Paradigms Workshop, 1998.
[32] M. Atallah, A. Elmagarmid, M. Ibrahim, E Bertino, and V. Verykios, "Disclosure limitation of sensitive rules", Workshop on Knowledge and Data Engineering Exchange, 1999.
[33] C. C. Aggarwal and P. S. Yu, "Privacy Preserving Data Mining:Models and Algorithms," 2007.
[34] L. Cao, H. Zhang, Y. Zhao, D. Luo, and Chengqi Zhang, "Combined Mining: Discovering Informative Knowledge in Complex Data," IEEE Transaction on Systems,Man, and Cybernetics:Part B, vol. 41, no. 3, June 2011.
[35] X. Yin, J. Han, J. Yang, and P. S. Yu, "Efficient classification across multiple database relations: A CrossMine approach," IEEE Trans. Knowl.Data Eng., vol. 18, no. 6, pp. 770–783, Jun. 2006.
[36] A.Holzinger, " On Knowledge Discovery And Interactive Intelligent Visualization of Biomedical Data," Research Unit Human-Computer Interaction, Institute for Medical Informatics, Statistics & Documentation, 2012.
[37] D. Zhanga,S.Chen, and Z.Zhou, "Constraint Score: A new filter method for feature selection with pairwise constraints," Journal of Pattern Recognition Society, October 2007.
[38] Y. Ji, H.Ying,P. Dews, A. Mansour, J. Tran,R. E. Miller, and R. Michael Massanari, "A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance," IEEE Transaction on Information Technology in Biomedicine, vol. 15, no. 3, May 2011.
[39] Goodwin Linda et al., "Data mining issues and opportunities for building nursing knowledge", International Journal of Biomedical Informatics, vol. 36, pp. 379–388, 2003.

All Rights Reserved, @IJAREST-2017

[40]M. D. Julie and B. Kannan, "Attribute reduction and missing value imputing with ANN: prediction of learning disabilities," Journal of Neural computing & application, vol. 21, pp. 1757-1763, May 2011.
[41] P. Merlin, A. Sorjamaa, B. Maillet, and A Lendasse, "X-SOM and L-SOM: A double classification approach for missing value imputation," Journal of Neurocomputing, vol.73, pp. 1103□1108, 2010.
[42] J. Luengo, J. A. Saez, and F. Herrera, "Missing data imputation for fuzzy rule-based classification systems," Journal of Soft Computing, vol.16, pp.863–881, October 2011.
[43] A. Zighed, S. Tsumoto, Z. W. Ras, "Mining Complex Data, Studies in Computational Intelligence," SCI.165, Springer-Verlag, pp.6-7, 2009.
[44] B. Zhu, C. He, and P. Liatsis, "A robust missing value imputation method for noisy data," Appl Intell. vol. 36, pp. 61–74, 2012.
[45] L. Xu, C.Jiang; J. Wang; J. Yuan, and Yong Ren, "Information Security in Big Data: Privacy and Data Mining," ,IEEE. Translations and content **mining**,vol.2,pp.1149,1176,2014
[46] M. Lin, K. Tang; X. Yao, "Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification," IEEE Transactions on Neural Networks and Learning Systems, vol.24, no.4, pp.647,660, April 2013
[47] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Trans. Syst., Man Cybern. B*, vol. 42, no. 4, pp.1119–1130, Apr. 2012.
[48] K. Sim, G. Yap; D. R.Hardoon, V. Gopalkrishnan, C. Gao ,and S. Lukman , "Centroid-Based Actionable 3D Subspace Clustering,", IEEE Transactions on Knowledge and Data Engineering, vol.25, no.6, pp.1213,1226, June 2013
[49] Y. Ren, G. Li, J. Zhang, and W. Zhou "Lazy Collaborative Filtering for Data Sets With Missing Values", *IEEE Trans. Syst., Man Cybern* Cybernetics, vol. 43, no. 6, pp. 1822 - 1834, Dec. 2013
[50] X. Liu, L.Wang, J. Zhang, J. Yin ,and H. Liu , "Global and Local Structure Preservation for feature Selection" ,IEEE Transactions on Neural Network and Learning Systems on vol.Z5 , pp. 1083 – 1095, 2014

All Rights Reserved, @IJAREST-2017