Subgroup Discovery a Data Mining Technique: Immense Survey

Deepali Nidhan Gunjate[1], **B. R. Kanawade**[2]

[1]Department of Computer, Dnyanganga College of Engineering and Research , Pune, India

[2]Dnyanganga College of Engineering and Research, Pune, India

Abstract: ***Subgroup discovery is a data mining procedure which concentrates fascinating rules regarding a target variable. A paramount feature for this task is the mixture of predictive and descriptive induction. An overview identified with the assignment of subgroup discovery is presented here. This survey concentrates on the establishments, algorithms, and progressed studies together with the applications of subgroup discovery. This paper shows a novel data mining systems for the investigation and extraction of learning from information created by electricity meters. In spite of the fact that a rich source of data for energy utilization analysis, power meters deliver a voluminous, quick paced, transient stream of information those traditional methodologies are not able to address altogether. So as to beat these issues, it is imperative for a data mining framework to consolidate usefulness for break summarization and incremental analysis utilizing intelligent procedures. In subgroups whose sizes are large and patterns are not usual has to be discovered. Their models have to be generated first. The many algorithms have been used to overcome the wider range of data mining problems. This paper gives a survey on subgroup discovery patterns from smart electricity meter data.***

Keywords: Subgroup discovery, data mining, pattern recognition, data analysis, knowledge discovery.

1. Introduction

The idea of subgroup discovery was at first presented by Kloesgen [1] and Wrobel[2], and all the more formally characterized by Siebes [3] yet utilizing the name Data Surveying for the discovery of subgroups. It can be described as [4]: In subgroup discovery, we consider we are given a in titled as people population (objects, client) and a property of those people we are interested in. The aim of subgroup discovery is then to find the subgroups of the population that are measurably "most interesting", i.e. are as huge as could be expected under the circumstances and have the most surprising statistical (distributional) attributes concerning the property of investment. Subgroup discovery try to find relations between distinct properties or variables of a set concerns with a target variable. Because of the truth that subgroup discovery is centered in the extraction of relations with interesting features, it is not important to acquire fully but partial relations. These relations are depicted as individual rules. The basic aim of this paper is to present an outline of subgroup discovery by examining the fundamental properties, models, quality measures and real issues understood by subgroup discovery approaches. Subgroup discovery [1,2] is an extensively pertinent data mining strategy focuses for finding intriguing relationship between distinctive objects in a set concerning a particular property which is of enthusiasm to the user the target variable. The patterns concentrated are ordinarily represented to as rules and called subgroups [3]. Previous methods have not possessed the capacity to attain this propose. For instance, predictive strategies increases exactness to effectively order new objects, and descriptive strategies basically scan for relations between unlabelled objects. The requirement for acquiring straightforward models with a higher state of interest prompted statistical strategies which hunt down irregular relations [1]. Thus, subgroup discovery is some place part of the way in the middle of supervised and unsupervised learning [5]. It can be viewed as that subgroup discovery lies between the extraction of association standards and the acquiring of classification guidelines.

Data mining is a phase of the Knowledge Discovery in Databases characterized as "the non-trivial extraction of unknown, implicit, and possibly helpful data from information" [6]. Explanation of the ten most utilized data mining algorithm can be found in [7]. Data mining systems can be connected from two alternate points of view:

- Predictive induction, in which, aim is the discovery of information for classification of prediction. Among its features, we can discover classification [8], regression [8], or temporal arrangement [9].
- Descriptive induction, in which fundamental target is the extraction of interesting learning from the information. Its features incorporate associationrules [10], summarisation [11] or subgroup discovery [1,2] can be specified.

The paper is organized as follows: Section II depicts literature review and overall study on subgroup discovery. Finally, concluding remark in Section III with future direction.

2. Related Work

On the study of the application of data mining in the field of electricity consumption and clients payment satisfaction and comparative related fields, Azadeh et al. [12], introduced an incorporated fuzzy framework and data mining methodology for the estimation of electricity capacity in Iran, they use decision tree and lookup table to concentrate the rule base which gives better solution, their careful investigation is focused around the aggregate electricityconsumption in Iran from 1992 to 2004, and the system they utilized ended up being a perfect option for fuzzy regression, particularly if data structure is evolving.

Pallegedara et al. [13], in the area of customer relationship management and clients risk analysis in the application of payment satisfaction in mobile correspondence, proposed a predictivedata mining model to diminish the rate of constrained mix (turn to subscribers) as a result of non-payment. The technique would describe open amounts for payer and non-payer subscribers, which will help keeping subscribers from overspending and eventually beating, their proposed model speaks to a modern however powerful model for information driven determination of credit limits.

In [14] Nizar et al., examined knowledge discovery as a part of databases (KDD) for load profiles of electricity client. His paper presents comparative analysis of the clustering systems used to focus on the load profiles of diverse electricity clients. In the study, a current load profiling systems are analysed utilizing data mining methods by analysing and assessing them. The target of his study is to focus the best load profiling strategies and data mining methodologies that group, recognize and predict non-technicallosses in the circulation area because of broken metering and charging slips, additionally to find learning on client behaviour and preferences to help in aggressive deregulated business sector.

At the point when applying subgroup discovery approaches, a few viewpoints must be considered. In this segment, we concentrate on depicting the suggestions identified with the pre-processing of the information and post-processing the knowledge, the discretisation of continuous variables, the utilization of domainknowledge, and the visualization of the results.

A. Scalability in subgroup revelation

At the point when applying data mining strategies to real world issues, these generally have high dimensionality, unavoidable for the majority of the typical algorithm. There are two ordinary potential outcomes at the point when a data mining algorithms does not work appropriately with high dimensional data sets [15]: upgrading the algorithm to run proficiently with huge data information sets or diminishing the size of the information without changing the result definitely.

Sampling is one of the systems most broadly utilized as a part of data mining to decrease the dimensionality of a data set and comprises of the determination of specific examples of the data set as indicated by some model. The application of a sampling method in the beginning database without considering conditions and relationship between variables could prompt a critical loss of information for the subgroup discovery task. On the off chance that it is important to apply some method to scaling down the data set in a subgroup discovery algorithm, it is particularly critical to guarantee that no imperative data for the extraction of interesting subgroups in the data is lost.

B. Pre-processing of the variables

It is exceptionally regular that a percentage of the variables gathered in the data sets used to apply subgroup discovery strategies are continuous variables. A large portion of the subgroup discovery algorithms are not ready to handle continuous variables. For this situation, a past discretisation can be connected utilizing distinctive mechanisms [16,17].

C. Domain knowledge in subgroup revelation

Utilizing domain knowledge as a part of data mining systems can enhance the nature of data mining results [18]. In subgroup discovery, it can help to find the centre on the subgroups identified with the target variable by confining the search space. Diverse the methodologies to incorporate the area information in subgroup discovery have been displayed as of late:

- In[19], the authors displayed the Semantic Subgroup Discovery in which semantically annotated learning sources are utilized as a part of the data mining process as background information. In this process, the results acquired have a complex structure which permits the specialists to see novel relationships in the information.
- In [20], Domain Knowledge is displayed as a "methodological approach for providing domain learning in a declarative manner”.

D. Different Applications of Subgroup Discovery

Our study of literature also focuses on various applications in which subgroup discovery has been used. Our plan is to decide load shading of electricity for this we are going to use Random forest algorithm. Following are different application areas in which subgroup discovery has been used.

1) **Subgroup discovery in the medical area**
   In the particular study, the accompanying gatherings of issues tended to by subgroup discovery can be discovered: detection of risk gatherings with coronary illness, brainischaemia, cervical cancer, and psychiatric crisis. Just about all the suggestions in the medicinal area were tackled through the DMS softwarewith the algorithm SD [21]. The psychiatric issue exhibited in [22] was solved with evolutionary fuzzy frameworks.
2) **Bio informatics issues solved through subgroup discovery**
   Distinctive genuine issues have been illuminated utilizing subgroup discovery as a part of the bio informatics space. These issues are portrayed by their high number of variables and low number of records in the databases. This makes it hard to concentrate interesting results.
3) **Subgroup discovery in e-learning**
   The configuration of online education frameworks has had a high development in the most recent years. These frameworks accumulate an incredible measure of profitable data when analysingstudent’s conduct then again distinguishing conceivable errors, deficiencies, and enhancements. Nonetheless, because of the enormous amounts of information, these frameworks can produce data mining tools which can help in this undertaking are requested [23].
4) **Spatial subgroup discovery**
   The combo of exploratory analysis of spatial data through geological visualization furthermore methods of subgroup discovery are talked about in distinctive papers in the particular book reference. The issues handled were identified with demographic [24].

3. Conclusion

An overview of research on subgroup discovery has been given in this paper; focus is to cover the early work in the field and in addition of recent work with the topic. The primary properties and components of this task have been introduced, and the all the more broadly utilized quality measures for subgroup discovery have been described for their properties. Likewise, distinctive applications of subgroup discovery methodologies to real world issues have been displayed, composed concerning the range of the application. This is a most emerging field, and there are a few open issues in subgroup discovery. A important issue to deliver is to figure out which quality measures are more adjusted both to assessing the subgroups discovered and to managing the search process. In the future work, we can find faulty meters which gives incorrect reading using pattern matching.

4. Acknowledgments

I would like to thank the GUIDE for drawing out attention to then-possible extent also the staff members and University of Pune. I also enjoyed the interesting and fruitful discussion with the friends working in this area. Finally I would like to thank my family and friends for their valuable support.

References

[1] Kloesgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances inKnowledge discovery and data mining. American Association for Artificial Intelligence, pp 249–271

[2] Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1stEuropean symposium on principles of data mining and knowledge discovery, vol 1263. Springer, LNAI,pp 78–87

[3] Siebes A (1995) Data Surveying: foundations of an inductive query language. In: Proceedings of the 1stinternational conference on knowledge discovery and data mining. AAAI Press, pp 269–274

[4] Wrobel S (2001) Inductive logic programming for knowledge discovery in databases. Springer, chapRelational Data Mining, pp 74–101

[5] Kralj-Novak P, Lavrac N, Webb GI (2009) Supervised descriptive rule discovery: a unifying survey ofconstrast set, emerging pateern and subgroup mining. J Mach Learn Res 10:377–403

[6] Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: anoverview. In: Advances in knowledge discovery and data mining. AAAI/MIT Press, pp 1–34

[7] Wu X, Kumar V, Ross-Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS,Zhou ZH, Steinbach M, Hand DJ, Steinberg D (2009) Top 10 algorithms in data mining. KnowlInfSyst14(1):1–37

[8] Cherkassky V, Mulier FM (2007) Learning from data: concepts, theory and methods, 2nd edn. IEEEPress, New York

[9] Box G, Jenkins G, Reinsel G (2008) Time series analysis: forecasting and control, 4th edn.Wiley, NewYork

[10] Agrawal R, Imieliski T, Swami A (1993) Mining association rules between sets of items in largedatabases. In: Proceedings of the 1993 ACM SIGMOD international conference on management ofdata. ACM Press, pp 207–216

[11] Zembowicz R, Zytkow JM (1996) From contingency tables to various forms of knowledge in databases.In: Advances in knowledge discovery and data mining. AAAI/MIT Press, pp 329–349

[12] Azadeh M.A., Ghaderi S.F., Guitiforooz A. and Saberi M., "Improved Estimation of Electricity Demand Function by Integration of Fuzzy System and Data Mining Approach", IEEE International Conference on Industrial Technology, ICIT 2006, pp.2160-2165, 15-17 Dec. 2006

[13] Pallegedara A., Amaratunga V.S., Gopura R.A.R.C. and Jayathileka P.D., "AI Based Approach of Predicting the Credit Limits of Users to Middle Customer based Mobile Communication Services", First International Conference on Industrial and Information Systems, pp.588-592, 8-11 Aug. 2006

[14] Nizar A.H., Dong Z.Y and Zhao J.H. , "Load profiling and data mining techniques in electricity deregulated market", Power Engineering Society General Meeting, IEEE pp.7, 2006

[15] Domingo C, Gavaldá R, Watanabe O (2002) Adaptive sampling methods for scaling up knowledgediscovery algorithms. Data Mining KnowlDiscov 6(2):131–152

[16] Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classificationlearning. In: 13th International joint conference on artificial intelligence, pp 1022–1029

[17] Liu H, Hussain F, Tan C, Dash M (2002) Discretization: an enabling technique. Data mining KnowlDiscov 6:393–423

[18] Richardson M, Domingos P (2003) Learning with knowledge from multiple experts. In: Proceedings ofthe 20th international conference on machine learning. AAAI Press, pp 624–631

[19] Lavrac N, Kralj-Novak P, Mozetic I, Podpecan V, Motaln H, Petek M, Gruder K (2009) Semanticsubgroup discovery: using ontologies in microarray data analysis. In: Proceedings of the 31stannual international conference of the IEEE engineering in medicine and biology society. IEEE Press,pp 5613–5616

[20] Atmueller M, Seipel D (2009) Using declarative specifications of domain knowledge for descriptivedata mining. In: Proceedings of the international conference on applications of declarative programmingand knowledge management and the workshop on logic programming, vol 5437. Springer, LNAI,pp 149–164

[21] Gamberger D, Lavrac N (2002) Expert-guided subgroup discovery: methodology and application. JArtifIntell Res 17:501–527

[22] Carmona CJ, González P, del Jesus MJ, Navío M, Jiménez L (2010b) Evolutionary fuzzy rule extractionfor subgroup discovery in a psychiatric emergency department. Soft Comput Special Issue on “GeneticFuzzy Systems” (in press)

[23] Romero C, Ventura S (2007) Educational data mining: a survey from 1995 to 2005. Expert SystAppl33(1):135–146

[24] Andrienko N, Andrienko G, Savinov A, Voss H,Wettschereck D (2001) Exploratory analysis of spatialdata using interactive maps and data mining. CartogrGeogrInfSci 28(3):151–165