Application of Uninorms to Market Basket Analysis

Raymond Moodley*

Centre for Computational Intelligence, De Montfort University, Leicester, UK

Email: raymond.moodley@my365.dmu.ac.uk, raymond.moodley@outlook.com

Francisco Chiclana, Fabio Caraffini

Centre for Computational Intelligence, De Montfort University, Leicester, UK

Email: chiclana@dmu.ac.uk; fabio.caraffini@dmu.ac.uk

Jenny Carter

University of Huddersfield, UK

Email: j.carter@hud.ac.uk

May 8, 2018

Abstract

The ability for grocery retailers to have a single view of customers across all their grocery purchases remains elusive and has become increasingly important in recent years (especially in the UK) where competition has intensified, shopping habits and demographics have changed and price sensitivity has increased following the 2008 recession. Numerous studies have been conducted on understanding independent items that are frequently bought together (association rule mining/ frequent itemsets) with several measures proposed to aggregate item support and rule confidence with varying levels of accuracy as these measures are highly context dependent. Uninorms were used as an alternative measure to aggregate support and confidence in analysing market basket data using the UK grocery retail sector as a case study. Experiments were conducted on consumer panel data with the aim of comparing the uninorm against three other popular measures (Jaccard, Cosine and Conviction). It was found that the uninorm outperformed other models on its adherence to the fundamental monotonicity property of support in market basket analysis. Future work will include the extension of this analysis to provide a generalised model for market basket analysis.

*Corresponding author

Keywords: Uninorms; Frequent itemset mining; Market basket analysis; aggregate measures; support; confidence

1 Introduction

The need to have an intimate understanding of the customer with a view of predicting their wants has always been a key ambition for retailers across the globe. This has become increasingly important in recent years as a result of increased competition and advances in technology that now makes this ambition more achievable.[14] The major UK grocery retailers have all embraced customer data analytics in an attempt to develop their customers, enhance loyalty and, consequently, sales. Given this, customer data analytics, in particular market basket analysis (MBA) and Associated Rule Mining (ARM), has become increasingly popular from a commercial perspective.[13]

ARM and frequent itemset mining first introduced in[1] is considered to be one of the most researched fields in data mining.[4] Central to this are the concepts of support (the ratio of the number of transactions containing any item X to the total number of transactions) and confidence (the ratio of the number of transaction containing both XY to the number of transactions containing X). Further, item X is said to be frequent if the number of transactions containing X is greater than a minimum support threshold (minsup) whilst items X and Y are considered to be associated if the confidence of (X leading to Y) or *vice versa* exceeds a minimum confidence threshold (minconf). Whilst it is obvious that the best XY targets for marketing by retailers are those that have the highest support and confidence, the situation is less clear when either parameter is high or low, for example: which is a better target? a XY combination that has high support and low confidence or a XY combination that has low support and high confidence? This problem is not new and has been widely studied over the years as it is considered to be at the heart of many data mining problems. Consequently several models exist for understanding the relationship between variables with some better suited to a specific task than others.[17]

It is against this backdrop that this paper proposes the use of the uninorm to model the relationship between support and confidence of items in a market basket. The UK grocery retail sector is used as a case study. The data used as part of the study is consumer panel data obtained from a third party analyst.[11] The proposed uninorm model is compared with the Jaccard, Cosine and Conviction models which are considered to be suitable for market basket analysis.[17] The benefits of this study include individual retailers enhancing their marketing activity leveraging the proposed model and their own retail-specific data. Armed with this data, individual retailers could engage in targeted marketing campaigns and capture the sales of products

that are being purchased elsewhere leading to increases sales and customer loyalty.

The rest of the paper is organised as follows. A review of related work is provided in Section 2. The problem statement and related mathematical model is presented in section 3 with the experimental methodology and results and discussion detailed in section 4 and section 5, respectively. The paper ends with conclusions, limitations and future work outlined in section 6.

2 Related work

2.1 Support and Confidence in Market Basket Analysis

The objective measures of support and confidence as outlined by[1] remain central to MBA with several studies re-affirming the downward closure property, $supp(A) \geq supp(AB)$, as key to effective modelling.[9] Bayardo *et al.*[3] extended this concept by exploiting monotonicity noting that whilst support is monotonically increasing with increased transaction volumes, that is $supp(A) \geq supp(B)$ if number of transactions containing A is greater than or equal to the number of transactions containing B, confidence behaves differently. Confidence is monotonically decreasing with respect to the support of the antecedent for a fixed support of the consequent. Hence if $supp(A) \geq supp(B)$ then $conf(A \rightarrow B) \leq conf(B \rightarrow A)$ for a fixed $supp(AB)$.

Minimum support and confidence constraints first introduced by[1] were born out of practical constraints, particularly in the retail sector where only the rules that passed a predefined level of interestingness were considered. However others[5, 10] noted that in some applications it is important to find all rules, including rare rules and hence setting minimum thresholds is not always easy or applicable.

2.2 Interestingness Measures in Market Basket Analysis

Combining variables to form interestingness measures has been well-studied over the years. Indeed, 38 such measures were noted in[8, 12] and 21 were compared in.[17] The primary reason for the plethora of measures is that the context plays a major role and whilst one measure may be best suited to one context, it may to be totally unsuitable to another.[17, 19] Given this, in their study of the 21 measures,Tan *et al.*[17] mapped the performance of the 21 measures against the contexts that they are best suited, showing that the cosine measure was best suited to retail settings.

A common thread across most measures was that they largely centred on Piatetsky and Shaprio's principles for rules interestingness.[8, 15, 17] The three principles are outlined as follows:

- **Statistical Independence**: Rules are not interesting if they are statistically independent, hence the measure (M) should be zero.
- **Monotonically Increasing**: The measure (M) increases with increasing *P(A,C)* for a fixed *P(A)* and *P(C)*. *P(A)*, *P(C)* and *P(A,C)* are defined using equation (3).
- **Monotonically Decreasing**: The measure (M) decreases for increasing *P(A)* whilst *P(A,C)* and *P(C)* remain fixed

2.3 Uninorms as Interestingness Measures

Uninorms, first introduced in,[18] have been widely used in aggregating variables for multi-criteria decision making. The primary reason for this is its versatility which is influenced by the neutral element that can be defined to best suit the context. In a study of customer satisfaction,[6] demonstrated that uninorms were superior to regression and correlated better with customer satisfaction theory. Similarly, uninorms have been shown to be effective in aggregating sentiments of experts in group decision making.[2]

A review of recent literature has shown no evidence of uninorms being applied to market basket analysis. This paper attempts to demonstrate the possible application of uninorms to market basket analysis.

3 Problem Statement

This section commences by outlining a set of definitions that will be used throughout this paper. Following this, the problem statement is laid out in three parts using a specific scenario of a generalised problem to both simplify and illustrate the key concepts.

3.1 Definitions

3.1.1 Items and Itemsets

Leveraging the definition in,[1] let $I = \{I_1, I_2, ..., I_m\}$, be a set of all items with binary attributes under the following assumptions:

- **Quality**: All brand variations of the same item I_i are of equal quality
- **Quantity**: All variation in package size is ignored and consumers purchase enough to last no more than one time period

Consumers purchase subsets of I, in the form of transactions, $T = I_2\ I_9\ I_{11}\ ..\ I_x$ etcetera. Transactions are detailed in a subsequent section. Itemsets are collections of items and are subsets of items in a transaction.

3.1.2 Users

A set of users, $U = \{U_1, U_2, U_3, ..., U_u\}$, make transactions and have the following properties:

- **Identifiable**: All users are identifiable to enable future targeted marketing campaigns
- **Household**: All users belong to a household, with size $f > 0$. Only one user in the household makes transactions.

3.1.3 Stores

All items are available in a set of stores, $S = \{S_1, S_2, S_3, ..., S_s\}$ and users are free-willed and make transactions in any number of stores contained in S. Stores have the following assumptions:

- **Confidentiality**: Stores do not share information or collude with each other
- **Collectivity**: All variations of the store (e.g. internet, convenience, superstore) are included under the same store.
- **Database**: Each store maintains a detailed, confidential, database of its transactions.

3.1.4 Time Period

Each transaction occurs in a single time period, where $W = \{W_1, W_2, W3, ...W_t\}$ is the set of all time periods under consideration. Time periods have the following assumption:

- **Collectivity**: Only one transaction per store, per user can occur in a time period

The *Collectivity* assumption takes into consideration the practical aspects of shopping and builds on the *Quantity* assumption outlined for items. The generally accepted length of a shopping period is one week as it is in line with how most people plan their household activity. Hence all items purchased within the week may be considered to be the market basket of the household.[7, 11, 16] Given this, it then becomes important to identify all items purchased in a given store in that week, hence the day or time of purchase is immaterial. Consequently all items purchased from the same store in the same week is aggregated into a single transaction. Having too short a time period is not practical for a sound analytical exercise as it will result in delinking of obvious shopping patterns. For example, if the period was one day then the model will suggest that items bought by the same user on Saturday in store S_1 are not related to items bought on Sunday in store S_2, for example. In practise this is not the case and consumers will generally spread their grocery shopping over the weekend or several days based on logistics.[11, 13] The assumption on *Quantity* is a necessary simplification and whilst it does impact frequently purchased item sets in that bulking buying of a single product in one

time period could make the overall frequent itemset less-frequent over several time periods, this behaviour is becoming more an exception that the norm given the changing shopping patterns of today's consumers.[7, 13]

3.1.5 Transactions

A transaction is a set of all items purchased and is recorded in a database. There are three dimensions to a transaction namely: store, user and period. Hence a transaction is fully detailed as $T_{s,u,t} = I_2\, I_5 \ldots I_i$ etcetera and may be read as "A unique transaction T occurring in store, s, for user, u, in time period, t, contains items I_2 I_5... etcetera". Note that the database is binary with the physical quantity of each item purchased being ignored. For simplicity the time period is usually a week and will be ignored, unless it is relevant.

A universal transaction database, D_0, represented by $S = 0$, is obtained by combining the transactions for all users in a given time period across all stores. Hence $T_{0,u,t} = T_{1,u,t} \cup T_{2,u,t} \cup \ldots \cup\ T_{s,u,t}$ etcetera is a combined transaction for user, u, and exists in D_0. This may be seen as the user's view of their transacting in a given period across all stores. In practice this universal database does not exist for all users, but large subsets exists through the consumer scanner panel programs run by third party analysts like[11] which are then used to make inferences of the overall market. It should be noted that $T_{0,u,t}$ is a binary transaction, consequently the same item I_i purchased from multiple stores in the same period by the same user will only be counted once as the database denotes 'presence' as opposed to quantity.

3.1.6 Support and Confidence

The standard definitions of support and confidence as outlined by[1] are included for completeness in equations (1) and (2).

$$\text{support of item, } \mathrm{I_i}|_{T_s} = supp(I_i)|_{T_s} = \frac{\text{Number of transactions containing } \mathrm{I_i}}{\text{Total number of transactions}}|_{T_s} \tag{1}$$

$$\begin{aligned}\text{confidence of item } \mathrm{I_i} \text{ leading to Items } \mathrm{I_i I_j}|_{T_s} &= conf(I_i \Rightarrow I_i I_j) \\ &= \frac{\text{Number of transactions containing } \mathrm{I_i} \text{ and } \mathrm{I_j}}{\text{Number of transactions containing } \mathrm{I_i}}|_{T_s}\end{aligned} \tag{2}$$

3.2 Problem Definition

Let r be a rule of the form $A \rightarrow C$ where $A = \{i_{A1}, i_{A2}, i_{A3}, \cdots, i_{Ak}\}; k \neq 0$ is a set of items extracted from I. Similarly $C = \{i_{C1}, i_{C2}, i_{C3}, \cdots, i_{Ap}\}; p \neq 0$ is a set of items extracted from I with $A \cap C = \emptyset$. The union

of A and C, $A \cup C$, is denoted as A, C. The support of A and the confidence of $A \rightarrow C$ is as follows:

$$\text{Support of A in T} = supp(A) = P(A) \tag{3}$$

Note that the support of C and A,C may be defined similarly using equation (3).

$$\text{Confidence of (A} \rightarrow \text{C) in T} = conf(A \rightarrow \text{C}) = \frac{P(A,C)}{P(A)} \tag{4}$$

In practice there are several such $A \rightarrow C$ combinations in the retail sector and it is the role of the marketing department of each retailer to decide on the optimum $A \rightarrow C$ combinations to target. Whilst it is obvious that items with both the highest support and confidence should be prioritised, it is less clear in cases where either the support or the confidence is high but not both.

As a simple, initial test of the appropriateness of a measure, we note from equation (4) that confidence is monotonically increasing with respect to the support of the consequent for a fixed antecedent. Hence, the appropriateness measure should be a monotonically increasing function as the support of the consequent increases. This has practical relevance as it demonstrates that item combinations with the highest support will be priority targets for marketing purposes.

The four appropriateness measures are detailed mathematically as follows:

$$\text{Cosine of (A} \rightarrow \text{C)} = \frac{P(A,C)}{\sqrt{P(A) \cdot P(C)}} \tag{5}$$

$$\text{Jaccard of (A} \rightarrow \text{C)} = \frac{P(A,C)}{P(A) + P(C) - P(A,C)} \tag{6}$$

$$\text{Conviction of (A} \rightarrow \text{C)} = \frac{1 - P(C)}{1 - conf(A \rightarrow \text{C})} \tag{7}$$

$$\text{Uninorm of (A} \rightarrow \text{C)} = \frac{P(A,C) \cdot conf(A \rightarrow \text{C})}{P(A,C) \cdot conf(A \rightarrow \text{C}) + (1 - P(A,C)) \cdot (1 - conf(A \rightarrow \text{C}))} \tag{8}$$

Real-life shopping data from the UK grocery retail will be used to establish consistent adherence to the monotonic principle by the four models outlined above.

4 Methodology

The 2012 annual consumer panel data set from Kantar was used to compare the effectiveness of the four models outlined in Section 3.2.The data consists of 32,000 users with over 51 million individual scanned items. The data was grouped into transactions using the principles outlined in Section 3.1. The confidence and support was calculated using a computer program written in R which leveraged the A rules module (based on the Apriori algorithm).[4] A minimum confidence of 0.12 and minimum support of 0.1 was specified to ensure that only stronger and frequent rules would be considered.

The algorithm generated an output for each of the twenty one stores and detailed the support and confidence for every (A $\rightarrow$ C) subject to the minimum support and confidence constraints.

5 Results and Discussion: Case Study Experiment

The results from the four models were plotted for several scenarios as outlined in figures 1 to 5. Based on the results, the uninorm correlates the best with support for the consequent given a fixed antecedent. The uninorm is always a monotonic increasing function whilst in most cases the conviction and cosines are not. The Jaccard coefficient is also monotonic and increasing in most cases but tends to fail when support and confidence of the rule is low and high respectively.

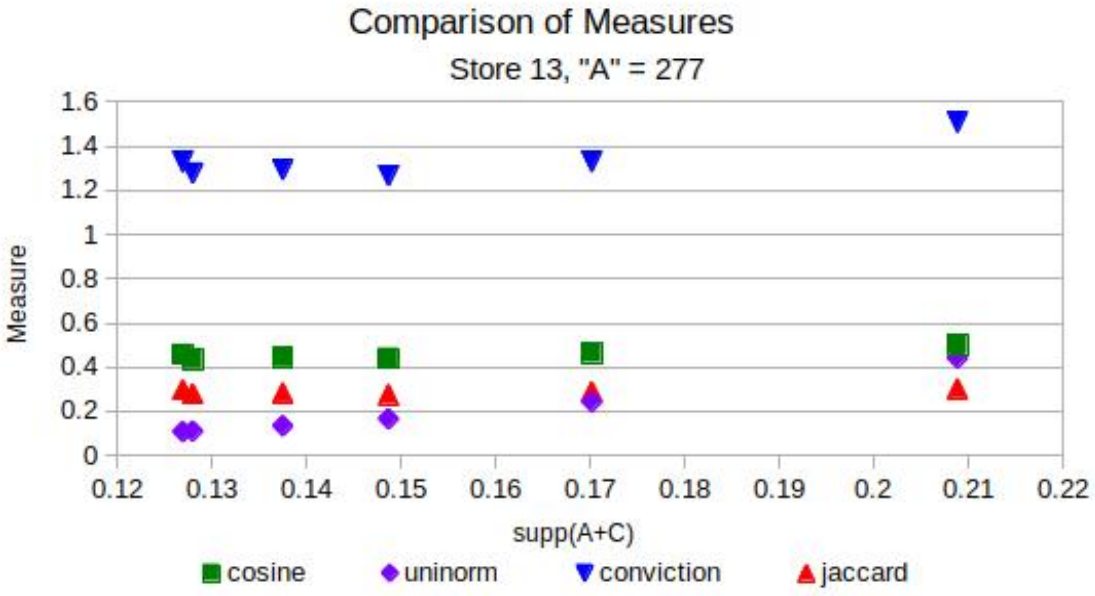

Figure 1: Store 13, A = Item 277

In figure 1 the data shows the performance of the four measures for the most popular item purchased in the UK grocery retail sector for a store that is not in the "big four" UK supermarkets. The "big four" is Tesco, ASDA, Sainsbury's and Morrisons. The results show that both conviction and cosine measures lacks monotonicity whilst the Jaccard and uninorm measures obey the monotonicity constraint.

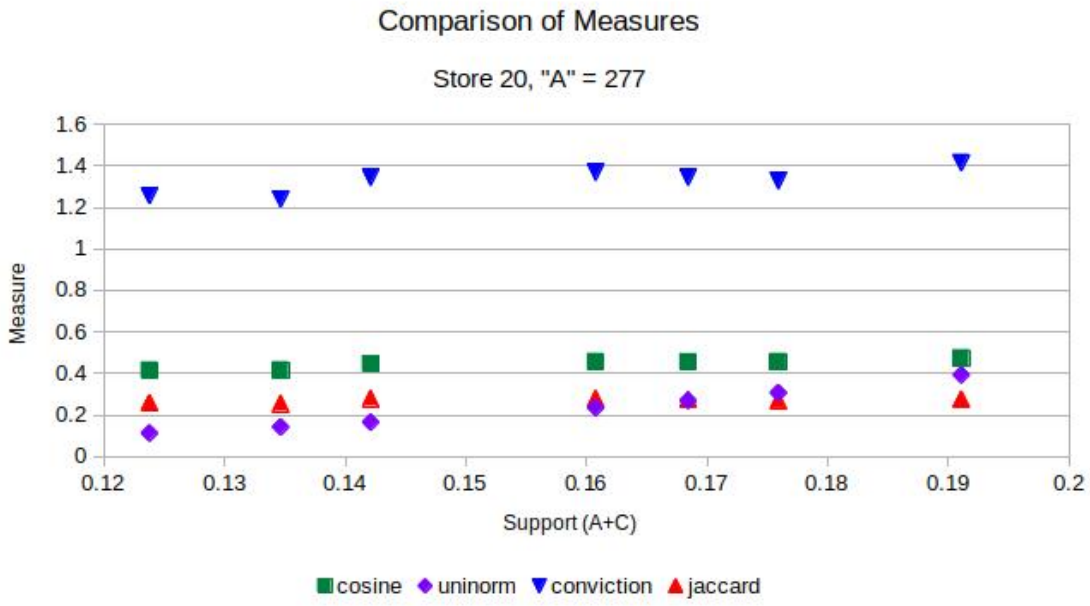

Figure 2: Store 20, A = Item 277

The results in figure 2 demonstrate a similar pattern for the most popular item purchased in a "big four" supermarket where volumes are considerably higher. Figure 3 depicts the scenario for a less-popular item at a store outside the "big four" where retail volumes are lower. Whilst the Jaccard, conviction and cosine measures all violate the monotonicity constraint, the uninorm remains consistent in obeying this constraint. A similar situation results for the same item purchased within the "big four" as shown in figure 4.

The Jaccard and uninorm measures were compared against each other using a "big four" supermarket and an item that has support close to the minimum support threshold (figure 5). From figure 5, it is clear that the uninorm is far superior in obeying the fundamental requirement of monotonicity with support.

6 Conclusion and Recommendations

Association rule mining is used widely particularly in the retail sector to perform market basket analysis. Several models exists to identify the best combination of items to target for marketing purposes, however these models varying in accuracy. In this study, the uninorm was used as alternative measure to identify the best combination of items for marketing purposes. A simple study involving the monotonicity property

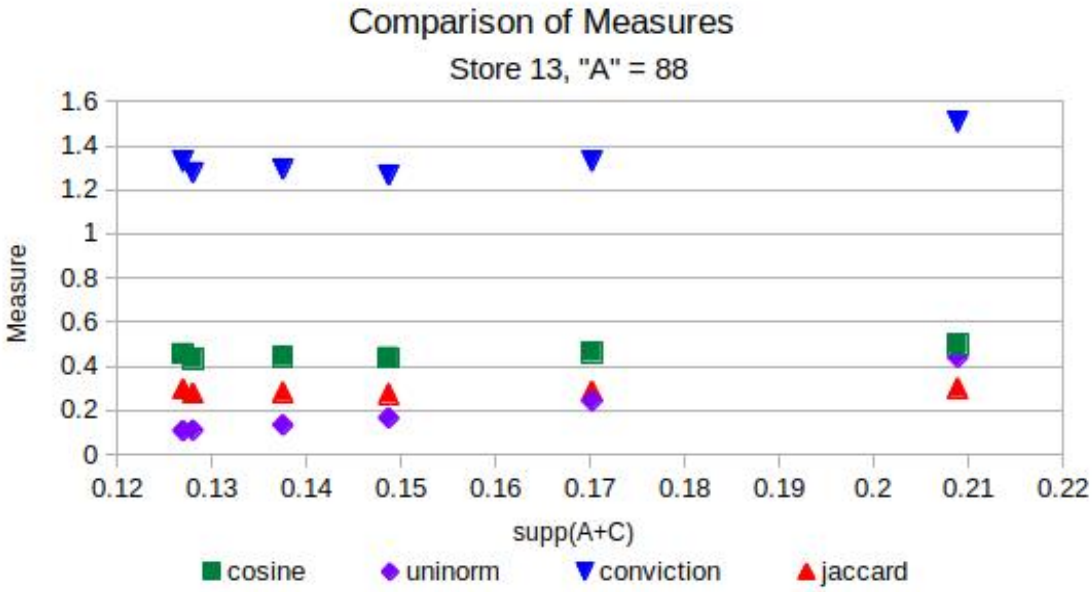

Figure 3: Store 13, A = Item 88

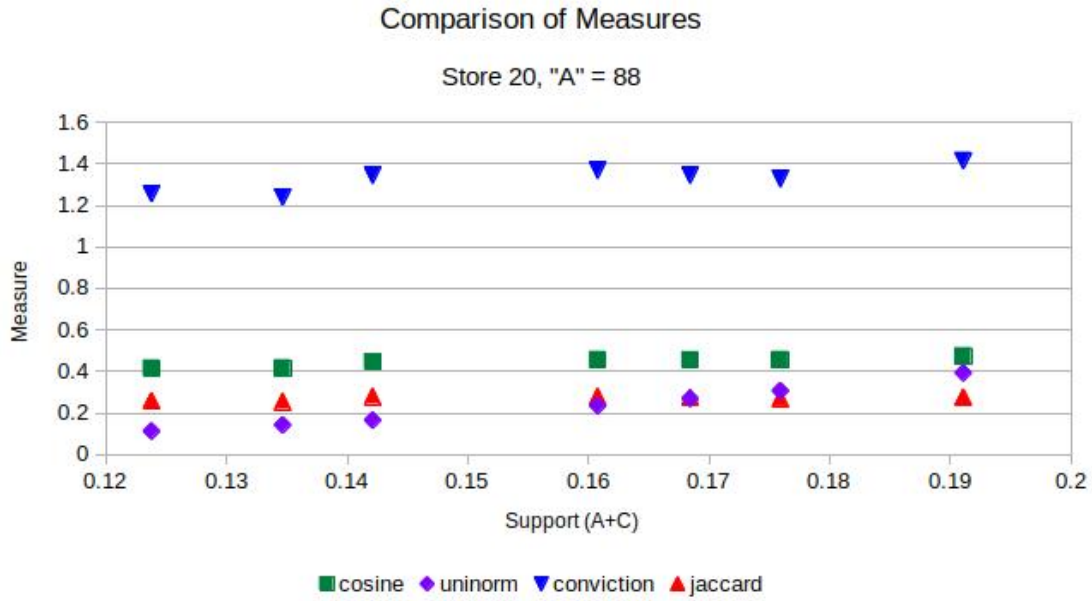

Figure 4: Store 20, A = Item 88

of market basket analysis was developed to test the performance of three popular models with the uninorm. The results showed that the cosine and conviction models were generally poor at obeying this property, whilst the Jaccard measure performed well when the antecedent has a high support but failed when the support was low. On the other hand the uninorm performed equally well irrespective of the support and consistently obeyed the monotonicity principle.

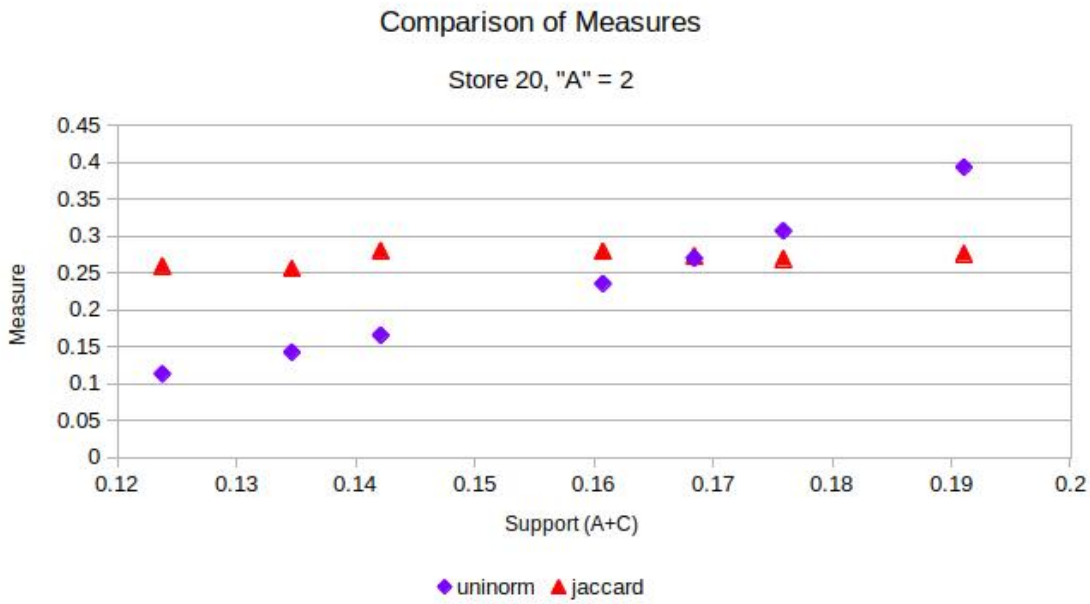

Figure 5: Store 20, A = Item 2

6.1 Limitations

We note that this is a preliminary study and the results are highly domain dependent. However the results demonstrates the adherence of the uninorm to a fundamental principle in market basket analysis. Our future work will look to expand on this and look to provide a more universal model for market basket analysis.

6.2 Future Work

Future work will include the use of the uninorm in developing a measure that will compare (A $\rightarrow$ C) and (B $\rightarrow$ D) combinations. Successfully developing a measure like this will enhance the ability of marketing departments to prioritise marketing initiatives across competing, independent combinations across the store.

References

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[2] Orestes Appel, Francisco Chiclana, Jenny Carter, and Hamido Fujita. Cross-ratio uninorms as an effective aggregation mechanism in sentiment analysis. *Knowledge-Based Systems*, 124:16–22, 2017.

[3] Roberto J Bayardo Jr and Rakesh Agrawal. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154. ACM, 1999.

[4] Christian Borgelt. Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):437–456, 2012.

[5] Toon Calders and Bart Goethals. Mining all non-derivable frequent itemsets. In *Principles of Data Mining and Knowledge Discovery*, pages 74–86. Springer, 2002.

[6] Benoit Depaire, Koen Vanhoof, and Geert Wets. The application of uninorms in importance-performance analysis. 2006.

[7] Fiona Ellis-Chadwick, Neil F Doherty, and Leonidas Anastasakis. E-strategy in the uk retail grocery sector: a resource-based analysis. *Managing Service Quality: An International Journal*, 17(6):702–727, 2007.

[8] Liqiang Geng and Howard J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.

[9] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[10] Jiawei Han, Micheline Kamber, and Jian Pei. Mining frequent patterns, associations, and correlations-6: Basic concepts and methods. 2012.

[11] Kantar. Kantar website. `http://www.uk.kantar.com/`, 2017. Accessed: 2017-01-30.

[12] Tien-Duy B Le and David Lo. Beyond support and confidence: Exploring interestingness measures for rule-based specification mining. In *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on*, pages 331–340. IEEE, 2015.

[13] Manthan. Top 3 grocery retail strategies using better shopper insights. `https://www.manthan.com/cpg-solutions/insights/503-top-3-grocery-retail-strategies-using-better-shopper-insights-1`, 2016. Accessed: 2017-01-30.

[14] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[15] Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, pages 229–238, 1991.

[16] Hongjai Rhee and David R Bell. The inter-store mobility of supermarket shoppers. *Journal of Retailing*, 78(4):225–237, 2002.

[17] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[18] Ronald R Yager and Alexander Rybalov. Uninorm aggregation operators. *Fuzzy sets and systems*, 80(1):111–120, 1996.

[19] T Yilmaz and A Guvenir. Analysis and presentation of interesting rules. In *Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks*, 2001.