Probabilistic Scoring of Validated Insights for Personal Health Services

Aki Härmä and Rim Helaoui
Philips Research
Eindhoven, The Netherlands

***Abstract*—In connected health services automatic discovery of recurring patterns and correlations, or insights, provides many interesting opportunities for the personalization of the services. In this paper the focus is on insight mining for a health coaching service. The basic idea in the proposed method is to generate a large number of insight candidates which have been pre-validated with domain experts and to score them using the data. The dynamic performance of the scoring is studied with a collection of lifestyle sensor data from volunteers. The proposed method is compared to a conventional data mining approach based on the *Apriori* algorithm. We demonstrate that the proposed method gives significantly more variability among the subjects and types of insights it finds which may reflect better the underlying statistics of individual lifestyle patterns of the different subjects.**

I. Introduction

Personal health coaching services typically focus on guiding the user to adopt a healthier lifestyle, for example, by being physically more active, sleeping and eating better. In the case of a conventional human health coach the opportunities for change are identified in a dialogue between the coach and the coachee. In an automated coaching machine based on a web service and an app, for example, it would be necessary to find those opportunities, or insights, automatically from the data [1].

Automatic generation of insights is a central topic in data mining literature. Conventional association mining is based on algorithms that find co-occurrences of sets of discrete data items [2], for example, particular books or food items in a marketing application, office behavioral data [3] or health data in clinical databases [4], [5]. In the case of health sensors with continuous data values, such associations cannot be uniquely defined but require a discretized and probabilistic framework for the description of insights [6], [7].

Let us call the proposed method the Probabilistic Scoring of Validated Insights, PSVI. The insights are found by computing probabilistic confidence scores for a large number of insight candidates which have been pre-validated by domain experts in the design phase. The pre-validation is necessary in a health coaching application to exclude potentially harmful insights. For example, the data may suggest that the user has a lower blood pressure on days when the user has slept less in the previous night. This insight may be interpreted by the user as an advice to sleep less while it most likely only refers to a correlation and not to a causal relation.

The proposed method can be seen as a modification of conventional association mining algorithms such as *Apriori* [8] or CHARM [9] but it has also interesting similarities with various machine learning algorithms. The proposed method also resembles recommendation systems [10], [11] but the problem is different and the same methodology is generally not applicable here.

In Sections II-III we give an overview of the PSVI algorithm and a use case in health programs. In Section IV the performance of the method is then studied using a collection of lifestyle sensor data from a group of volunteers, and the final results are discussed in Section V.

II. PSVI algorithm

Conventional association rule mining is based on counting co-occurrences of discrete items $\{I_k, I_j\}$ [12], [2]. In case that the items are continuous measurements such as a step count and heart rate the algorithm can also be applied after discretization [7]. This is often performed by dividing the continuous measurement range into a small number of bins and using for example fuzzy membership functions to describe the associations [13]. The presented PSVI method uses a relative discretization where one measurement in a context is either smaller or larger than another measurement. In the *Apriori* algorithm the discovery of insights would be then based on occurrences of the these cases. This basic method is developed and tested further in this document and it is shown that it is not necessarily efficient for dynamic selection of *interesting* insights.

The *interestingness* score of an associative rule can be characterized in many different ways, see, e.g., [14]. Due to the probabilistic nature of the insights discussed in this paper let us call this score a *confidence* value of an insight. For the purposes of this paper a high confidence should be related to a detection of an opportunity, which is typically a context or condition that somehow stands out from the data. For example, an insight may state that "a user walks less on Mondays than on Tuesdays". The confidence value of this statement should be based on (1) the statistics of walking on those weekdays, e.g., based on data from an activity bracelet, and (2) the observation that it differs from some other context.

In an insight mining application we might be interested in finding the highest scoring insight out of a collection of statements of the following form "in context *a* you walk less than in context *b*", where *a* and *b* could be, for example, two

different weekdays. Thus, a collection of statements can be defined as a set s_N of N triples $\{A_n, B_n, M_n\}$, i.e., $s_N = \{\{A_1, B_1, M_1\}...\{A_N, B_N, M_N\}\}$, where each triple refers to a particular statement or a measurement M_n and a pair of contexts A_n and B_n. Let us denote by $Pr(X_{A_n}|A_n)$ the conditional probability distribution of the measurement values X_{A_n} of the n^{th} statement conditioned on the context A_n, and by $p_n(x_{A_n})$ its probability density function (PDF). For example, if the n^{th} statement is "on Mondays you walk less than on Tuesdays", then $p_n(x_{monday})$ refers the PDF of the conditional probability distribution $Pr(StepCount|Weekday = Monday)$ and $p_n(x_{Tuesday})$ to that of the conditional probability distribution $Pr(StepCount|Weekday = tuesday)$.

The difference between two probability distributions defined over the n^{th} statement can be characterized by a divergence measure d_n, given by

$$d_n = D(p_n(x_{A_n}), p_n(x_{B_n})). \tag{1}$$

Typical divergence measures D are the Kullback-Leibler or Hellinger divergence. The divergence measures give the value 0.0 if the PDFs are identical and 1.0, if they do not overlap. Based on this, we generalize the diversion measure to the cases where x_{A_n} or / and x_{B_n} are scalars. Example statements of these cases would be "yesterday you walked more than on a typical Monday", and "today you walked more than yesterday", respectively. For the former case we first normalize the PDF in question to give 1.0 at its maximum and denote the normalized variant by $\bar{p}_n(x) = p_n(x)/\max(p_n(x))$. The divergence measure is then defined as follows.

$$d_n = 1 - \bar{p}_n(x_{A_n}|x_{B_n}) \tag{2}$$

Finally, for the case of comparing two discrete variables x_{A_n} and x_{B_n}, we calculate the divergence as indicated in Equation (3), where d_m is a normalization constant reflecting the range of interesting measurement values.

$$d_n = 1 - \exp(-\nu(x_{A_n} - x_{B_n})^2/d_m) \tag{3}$$

Combining the three metrics introduced above (Equations (1, 2) and (3)) we define the *PSVI divergence* measure for the n^{th} statement as follows.

$$\text{PSVI } D_g = \begin{cases} d_n, \text{if both } x_A \text{ and } x_B \text{ are distributions} \\ 1 - \bar{p}_n(x_{A_n}|x_{B_n}), x_A \text{ or } x_B \text{ is scalar} \\ 1 - \exp(-\nu(x_{A_n} - x_{B_n})^2/d_m), \text{scalars} \end{cases} \tag{4}$$

In the last part d_m is a scaling factor for different measurements, and in the following experiments the coefficient ν in the exponential was set to $\nu = 42$. Next to a divergence measure it is also necessary to take into account other factors that influence the confidence value such as the amount and quality of data used to score the statement. For that purpose, we include an additional term $W_c = 1 - \exp(-c/\alpha + \gamma)/\beta$ which adds a penalty to the confidence score in case the count c of the measurements in the context a or b is low. In Section IV the coefficients were set to $\alpha = 0.8$, $\beta = 2.0$, and $\gamma = 1.3$ which were found experimentally.

For a collection of N statements the insight with the highest confidence is then given by

$$s_w = \text{argmax}_{s_n} W_c D_g(s_n) \tag{5}$$

Equation 5 resembles in some sense a Bayesian classifier which selects the class (insight candidate) giving the highest likelihood (confidence). However, in PSVI the scoring model is typically not trained using statistical learning methods and the number of classes can be very large compared to a common multi-class classification case.

III. Use case in a personal health service

Let us say that the goal is that the service would provide a new personal insight about the behavior of the user, for example, every day. This requires a large number of insight candidates. Moreover, let us assume that each insight has a fixed probability p_w to score above some threshold. Based on the binomial distribution, the probability that all N cards are below the threshold on a given day is $(1 - p_w)^N$. Thus, in order to have at least one insight above the threshold, there should be

$$N > \log(1 - p_\gamma)/\log(1 - p_w) \tag{6}$$

insights to choose from. For example, having a $p_w = 0.05$, one needs more than 100 insight candidates to have $p_\gamma = 0.99$ probability that at least one of them scores above the threshold on one day. It is often desired to avoide repetition and show each insight only once, which would give a practical rule of thumb that the number of candidates should be at least 100 times the total number of insights the user is shown in the course of the health service.

The interface to a personal health service is often a smartphone app connected to a wearable device such as an activity bracelet. The app provides education and motivation that helps the user to change their behavior to healthier and feedback on the positive achievements. A typical activity bracelet produces a few measurement variables, for example, on step counts, energy expenditure, and heart rate. In addition, an app associated with the bracelet may track location data which makes it possible to detect when the person is at home, work, or outdoors, for example, and divide a day into segments such as morning or afternoon.

The conditional PDFs of all the measurement variables can be estimated for each segment separately and the conditions can represent, for example, different weekdays or other conditions in data selection. By different combinations of measurements, segments and conditions one can create a large number of comparative statements of the type "in context *a* and segment *c* your measurement M is typically higher than in context *b* and segment *d*". The content can be generated using Natural Language Generation, NLG, tools available for example in [15]. The combinatorics leads easily to hundreds of thousands of insight candidates.

Family	Nr	Example
I	3023	On Thursday morning your step count is K% lower than on Tuesday
II	1455	On Sunday evenings you burn less calories than on other weekdays
III	783	On Fridays your sedentary time is lower than the community average

TABLE I
EXAMPLES FROM THE THREE INSIGHT FAMILIES

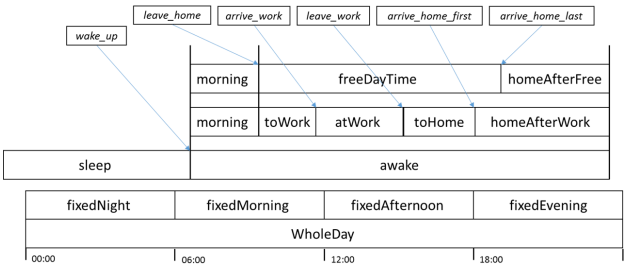

Fig. 1. Segmentation diagram. Change points between segments are typically found from GPS location data collected with an app.

In the experiments reported in this paper we use three families of insights, listed in Table I, with the total of 5261 *validated* insights. By validation we mean that the insights candidates were selected by eliminating combinatorical options that have no value for the application. For example, an insight that the user "walks more on Tuesday afternoon than on Friday morning" was eliminated but a similar insight comparing Tuesday and Friday afternoons was considered useful. The insight family *I* consists of direct comparisons of a measurement value in two contexts, e.g., Monday and Tuesday morning in one individual user. The family *II* compares a measurement in a particular context to a complement such as Monday v.s. other weekdays. Finally, the *III* family compares the data from one user to the averaged PDFs of the entire user community.

The conditional probabilities of the measurement values were estimated in a number of different day types and parts of a day. In the current paper there are eight activity-related measurement variables such as step count, active minutes, sedentary time, and maximum heart rate. The contextual setting contain separate weekdays, and combined statistics of all days, all work days, and all free days. In addition, the scalar measurements of "today" and "yesterday" were also included. The daily data was further segmented according to the diagram shown in Fig. 1. For example, the conditional probability distribution $Pr(X_{A_n}|A_n)$ introduced above could, for example, represent the distribution of step counts (X_{A_n}) in the morning segment on the condition that they day is Monday (A_n). Note that the total combinatorics of the contexts, measurements, and directions in this case would lead to more than a half million insight candidates.

The conditional probability distributions were based on the normal distribution model, i.e., the mean values and standard deviations. One practical reason is that the normal distribution has a clear concept of the mean which is needed in the communication of the insight to the end user. For example, in the insight: "on Mondays you are more active than on Tuesdays", the comparison refers to an average Monday and Tuesday and the confidence score is based on the divergence between the PDFs defined on the measurement values on Monday and Tuesday, respectively. The use of a normal distribution also leads to a very efficient computation of divergence metrics compared to other alternatives and the practical experience is that the long-term activity data is often predominantly normally distributed.

In the insight collection there are typically pairs of insights, e.g., in "In A, M is larger than in B" and "In A, M is smaller than in B". Respectively, each insight is associated with a sign attribute $\Sigma_n = -1$, or 1, to indicate the direction of the insight. The scoring of a collection of N insights can be performed using the following pseudo-code

0 $n = 0$
1 Compute the confidence score for $C_n = D_g\{A_n, B_n, M_n\}$
2 if $(E[Pr(X_{A_n}|A_n)] - E[Pr(X_{B_n}|B_n)]) \neq \Sigma_n$, then $C_n = 0$
3 $n = n + 1$ and return to Step 1 unless $n = N$

IV. EXPERIMENTAL RESULTS

The experimental data consists of lifestyle sensor data from 17 volunteers, health office workers with regular daily and weekly patterns. The subjects were using a smartphone app that records the location and activity data (moves-app) and a wrist-worn activity sensor which collect step counts, activity information, and heart rate measurements. In order to study the properties of the PSVI scoring mechanism in the population the insight candidates were scored on different days from the start of the data collection to the end date, which was 20-60 days after the start date. The number of insight candidates that score above 60% of the confidence value range in the three families as a function of time are shown in Fig. 2. The conditional PDFs were computed starting from the first day. In the first days only insights that compare the measurements of the current day to the community score high. The number of available insights grows in the following days and stabilize within approximately three weeks. Fig. 3 shows the number of subjects where a particular insight (x-axis) scores above 80% after 33 days of data collection. Approximately 14% of the cards did not score high in any of the 17 users and the number of cards that scored high for multiple users was low. The average probability that an insight scores above 80% is 2.5%. Equation (6) would suggests, although assuming a uniform probabilities, that there should be at least 178 insights for each selection to be 99% sure that there is at least one insight for every day.

In a typical application of PSVI the insights that have already been shown to the user would be removed from the collection and therefore the number of available insights would reduce over time. The simulation of the insight counts for all users is shown in Fig.4. It seems that in this collection of

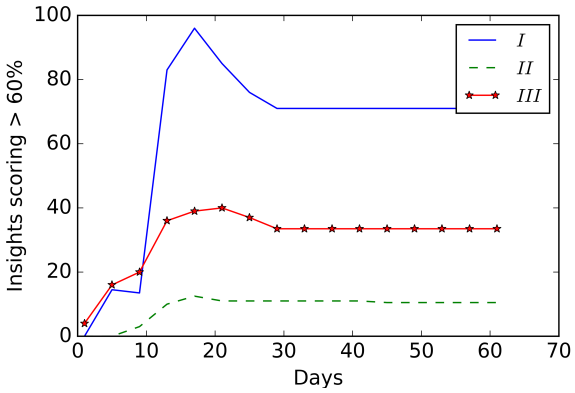

Fig. 2. The number of insight candidates scoring above 60% in the three families.

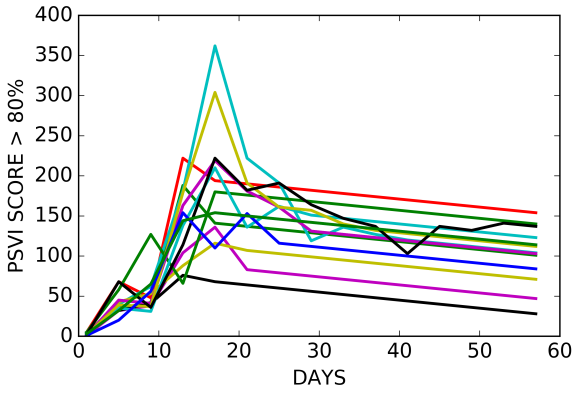

Fig. 4. The number of insight cards above 80% as a function of time in a content feed simulation where the top scoring insight is shown once per day in the program and removed from the collection.

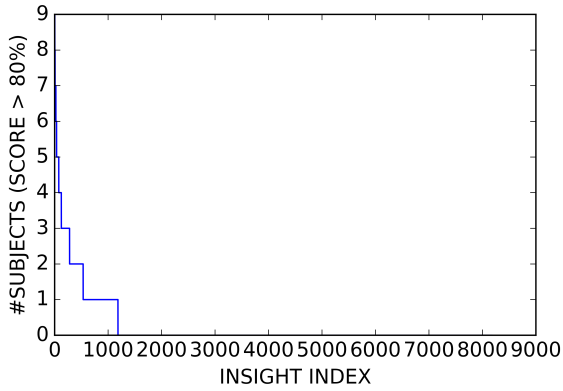

Fig. 3. The number of subjects where an insight on a sorted list of insights scores above 80% after 33 days.

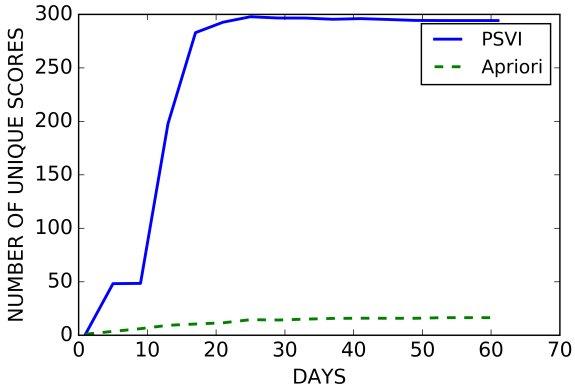

Fig. 6. The number of unique score values.

5261 validated insight is sufficient for at least 2-3 months of daily non-repeating insight messages for each of the 17 users, which is larger than the minimal value $5261/178 \approx 30$ days suggested by the rule of Equation (6).

A. Comparison of Apriori and PSVI scoring methods

To compare the popular Apriori and the PSVI scoring methods, the Apriori method was adapted for the selection of insights form a pre-defined collection. In practice, we computed the number of occurrences of the conditions related to the pre-scripted statement, for example, the number of times step counts per minute on Monday morning are higher than on Monday afternoon. The count was divided by the total count of Mondays in the data series to produce an Apriori interestingness measure. This was performed only for the 3023 insights from the family *I* of Table I because there is no unique way to count the corresponding differences in the families *II* and *III*.

The confidence values for each insight in the two methods after 9 days, and 51 days of data are shown in sub-panels of Fig. 5. The number of insights with a non-zero confidence score increases in going from 9 days to 51 days of data in both methods. However, the Apriori method gives a large number of insights with the same score. This difference is more pronounced in Fig. 6 which shows the number of unique score values in the collection of 3023 insights as a function of time from the start of the program. In the first days, the number of unique confidence values is similar in the first days but thereafter the proposed PSVI method has a significantly larger variability in insight scoring while with the Apriori scoring method hundreds of insights get the same confidence value.

The variability in the confidence scores suggests, but does not prove that the insights based on the PSVI method would be more *individualized* than the ones from the Apriori method. The mean inter-subjective correlation coefficients between the confidence values between all subjects in Fig. 7 may be considered as a more direct evidence of the personalization of the insight content. In the Apriori method the correlation value between subjects is significantly higher than in the proposed method and increases when the data grows, while in the PSVI method the correlation between confidence values reduces. One may expect that the differences between subjects become more clear when the mining algorithm has more data available about every subject, however, without ground truth this can be

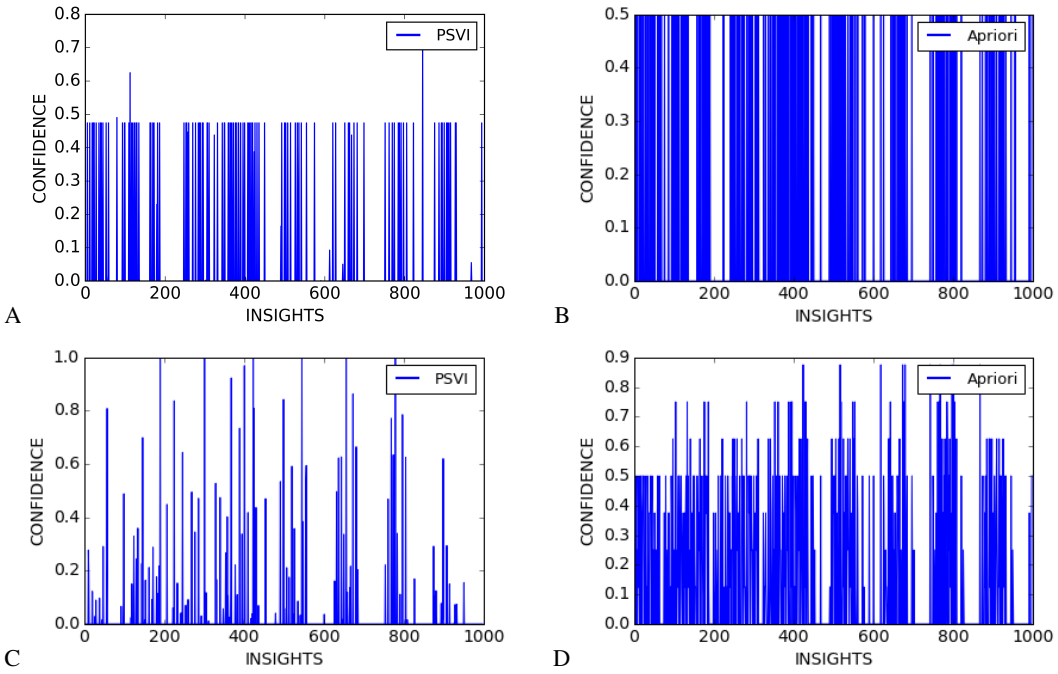

Fig. 5. Confidence spectra of the 3023 insight candidates in the two methods after 9 days of data (left) and 51 days of data (right) in one test subject. The insight candidate with the highest score is shown on the title line.

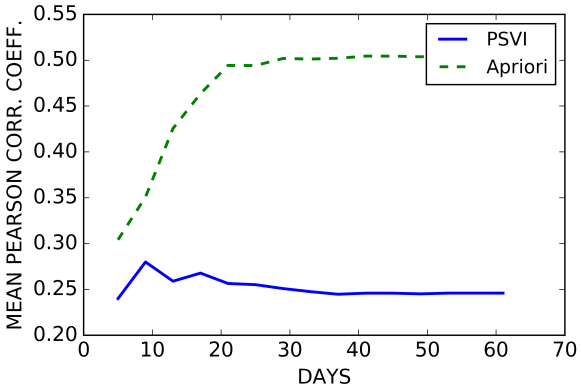

Fig. 7. Average Pearson correlation coef .

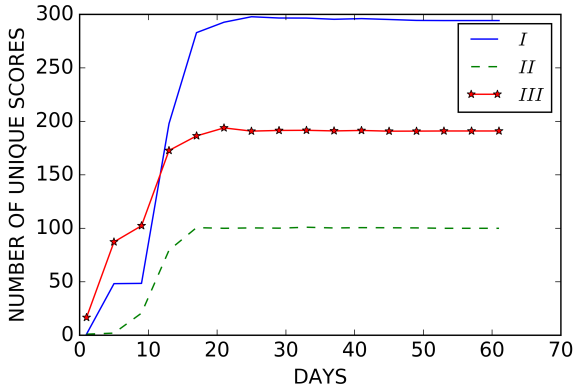

Fig. 8. The number of unique PSVI score values in the three families as a function of time.

only assumed to indicate a higher level of personalization.

B. Inter-subjective correlations in insight families

The temporal dynamics in the three insight families are illustrated in Figs. 8 and 9. The three insight families have similar temporal dynamics in counts of unique insights and in the reduction of intersubjective correlations over time.

V. Discussion and Conclusions

Classic data mining methods such as Apriori are, in principle at least, unsupervised methods that discover recurring associations between data items. The method (PSVI) introduced in this paper is based on probabilistic scoring of a large collection of pre-validated insight candidates. The method can be seen as an example of a *supervised* discovery method because all insights are known in advance. In practice the difference is very small if the pre-validated collection contains the entire space spanned by the data. The pre-validation is a necessary element to eliminate nonsense and potentially harmful insights in a health application. However, it also requires that all data items are known in advance. This is typically the case for example in a health service applicaiton but not the case, for example, in conventional data mining applications for example in online marketing where new items are continuously appearing.

In the paper we have introduced a new method, PSVI, for automatic discovery of insights from continuous multivariate

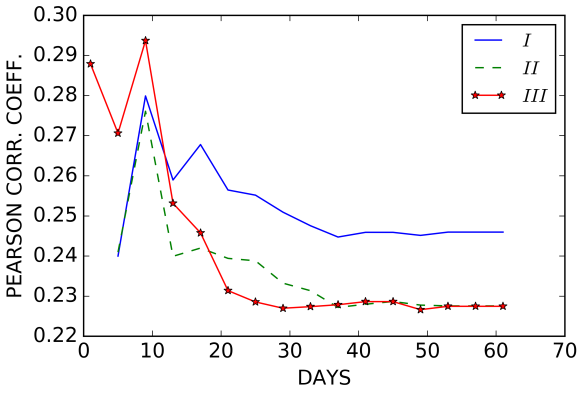

Fig. 9. The intersubjective correlation coefficient in the three insight families as a function of time.

time-series. The method was compared to a conventional data mining approach based on an adaptation of the famous Apriori algorithm [8] to the same use case. The results suggest that the PSVI provides more variability in the scoring of insights between individuals and a better resolution in the score values than the Apriori method based on counting associations. This may indicate that the proposed method provides insights that are more personal than the insights provided by Apriori. However, it should be noted that the Apriori method based on counting frequencies of co-occurrances is not well-suited for the current application where the differences between data items are in distributions of measurements rather than in frequencies of conditions.

The design of the insight library is a critical step and the analysis and experiments given in the paper give some guidelines on how it should be designed to meet the requirements on confidence and availability of insights. In particular, a simple rule of thumb in Equation (6) seems to be appropriate for the selection of the minimal number of insight candidates.

In the current paper only numeric evidence of the performance was reported. The current authors already have encouraging initial user test results from a small user panel in the application of an automated health service but a proper testing is a part of future work. The PSVI method can be used in all applications where there is a need for a controlled and pre-validated mining of comparative insights from sensor. The performance numbers depend on platform but one may anticiapte the proposed scoring method is computationally somewhat more expensive than the Apriori algorithm. The method has been implemented in a cloud-based health data service platform [16] where the PSVI, implemented using the Apache Spark library primitives [17], can score approximately 20000 insights per second and per processing node.

REFERENCES

[1] H. op den Akker, M. Cabrita, R. op den Akker, V. M. Jones, and H. J. Hermens, "Tailored Motivational Message Generation," *J. of Biomedical Informatics*, vol. 55, no. C, pp. 104–115, Jun. 2015.

[2] R. Agrawal and J. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 962–969, Dec. 1996.

[3] S. J. OMalley, R. T. Smith, and B. H. Thomas, "Data Mining Office Behavioural Information from Simple Sensors," in *Proceedings of the Thirteenth Australasian User Interface Conference-Volume 126*, Melbourne, Australia, Jan. 2012.

[4] S. Stilou, P. D. Bamidis, N. Maglaveras, and C. Pappas, "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare," *Studies in Health Technology and Informatics*, vol. 84, no. Pt 2, pp. 1399–1403, 2001.

[5] G. H. Gonzalez, T. Tahsin, B. C. Goodale, A. C. Greene, and C. S. Greene, "Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery," *Briefings in Bioinformatics*, vol. 17, no. 1, pp. 33–42, Jan. 2016.

[6] Y. Aumann and Y. Lindell, "A Statistical Theory for Quantitative Association Rules," *Journal of Intelligent Information Systems*, vol. 20, no. 3, pp. 255–283, May 2003.

[7] S. Born and L. Schmidt-Thieme, "Optimal Discretization of Quantitative Attributes for Association Rules," in *Classification, Clustering, and Data Mining Applications*, ser. Studies in Classification, Data Analysis, and Knowledge Organisation, D. D. Banks, D. F. R. McMorris, D. P. Arabie, and P. D. W. Gaul, Eds. Springer Berlin Heidelberg, 2004, pp. 287–296, dOI: 10.1007/978-3-642-17103-1_28.

[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. of 20th Intl. Conf. on VLDB*, 1994, pp. 487–499.

[9] M. J. Zaki and C. J. Hsiao, "Efficient algorithms for mining closed itemsets and their lattice structure," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 462–478, Apr. 2005.

[10] P. Resnick and H. R. Varian, "Recommender Systems," *Commun. ACM*, vol. 40, no. 3, pp. 56–58, Mar. 1997.

[11] M. Wiesner and D. Pfeifer, "Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges," *International Journal of Environmental Research and Public Health*, vol. 11, no. 3, pp. 2580–2607, Mar. 2014.

[12] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," in *Proc. 1993 ACM SIGMOD Int. Conf. Management of Data*, Washington DC, USA, 1993, pp. 207–216.

[13] H. Zheng, J. He, G. Huang, and Y. Zhang, "Optimized fuzzy association rule mining for quantitative data," in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Jul. 2014, pp. 396–403.

[14] A. A. Freitas, "On rule interestingness measures," *Knowledge-Based Systems*, vol. 12, no. 56, pp. 309–315, Oct. 1999.

[15] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. Beijing ; Cambridge Mass.: O'Reilly Media, Jul. 2009.

[16] F. Andry, R. Ridolfo, and J. Huffman, "Migrating Healthcare Applications to the Cloud through Containerization and Service Brokering:," in *HEALTHINF 2015*. SCITEPRESS - Science and and Technology Publications, 2015, pp. 164–171.

[17] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster Computing with Working Sets," in *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 10–10.